



UNIVERSITÉ DE NANTES



IAE NANTES
ÉCONOMIE & MANAGEMENT

Master 1 Économétrie et Statistiques, parcours Économétrie Appliquée

Étude sur le taux de pauvreté en 2015

Analyse de données sur 59 pays à travers le monde

Auteurs :

Nicolas DAVOINE

Diane THIERRY

Année universitaire 2019-2020

Sommaire :

Partie 1 : Introduction	2
Partie 2 : Analyse détaillée du sujet	5
Partie 3 : Analyse en Composantes Principales (ACP)	23
Partie 4 : Corrélation entre la variable à expliquer et les variables latentes	46
Partie 5 : Conclusion et discussion des résultats	56
Partie 6 : Annexes	57
Partie 7 : Bibliographie	68
Partie 8 : Table des matières	70

Partie 1 : Introduction

Dans cette introduction nous allons nous concentrer sur l'évolution historique de la pauvreté dans le monde et ses différents aspects afin de contextualiser notre analyse de données.

La pauvreté est un fléau mondial principalement alimenté par les crises économiques, financières mais aussi historiques ou encore climatiques. Les scientifiques, les décideurs politiques et les professionnels du développement s'accordent sur l'aspect "multidimensionnel" de la pauvreté. Avant la première moitié du XXème siècle la grande majorité de la population mondiale vivait dans la pauvreté ; seuls les pays d'Europe de l'Ouest et d'Amérique du Nord étaient suffisamment développés pour répondre aux besoins primaires de la majeure partie de leur population. Ce n'est qu'à la fin de la seconde Guerre Mondiale que la plupart des pays de la planète coopèrent afin de tenter d'éradiquer à long terme la pauvreté dans le monde.

Influencées par l'expérience du plan Marshall Américain (1948-1952) et par la signature des accords de Bretton Woods en 1944, de nombreuses institutions financières internationales ont été créées telles que la Banque Mondiale, le Fond Monétaire International (FMI) ou encore l'Association Internationale de Développement (IDA). Parmi les objectifs de ces institutions, le plus important est celui d'aider les pays considérés comme 'pauvres' à amorcer leur développement économique pour réduire la misère dans ces pays, afin de maintenir une paix durable à l'échelle mondiale. En effet, l'Aide Publique au Développement (APD) s'établit dès les années 1960 comme un outil de coopération internationale.¹

De ce fait ces organismes ont accordé des prêts à long terme aux pays les plus pauvres de la planète. Cela a permis de financer des programmes aidant à (re)construire les politiques, les institutions, les infrastructures et le capital humain nécessaires à un développement à la fois équitable et soutenable. Voici quelques exemples de projets de développement réalisés par l'APD dans les pays bénéficiaires : aides pour l'éducation, la santé, les dispositifs de sécurité sociale, l'approvisionnement en eau potable, les administrations publiques, les

¹ <http://www.sciencespo.fr/ceri/fr/carto/dossier-carto-la-pauvrete-dans-le-monde> (consulté le 24/09/2019)

infrastructures, l'agriculture et le développement rural. Ces interventions ont contribué à faire baisser la pauvreté et l'extrême pauvreté dans le monde².

Depuis les années 1950, le nombre de ménages considérés comme pauvres a diminué également grâce aux vagues successives d'immigration de populations des pays du Sud allant vers le Nord pour bénéficier des meilleures conditions de vie. L'**immigration** désigne l'entrée de personnes étrangères dans un pays ou une aire géographique pour y séjourner quelques temps ou pour s'y installer³. Ainsi, les ménages fuyant la misère de leur pays d'origine ont pu avoir accès aux conditions de vie des pays d'accueils (généralement les Etats-Unis et les pays d'Europe de l'Ouest). De ce fait, le nombre de personnes vivant dans la pauvreté s'est réduit dans les pays du Sud, sans forcément augmenter dans les pays développés puisque leurs économies de marchés permettaient d'employer aussi les nouveaux entrants.

Les répercussions économiques d'une telle vague d'immigration sont fortes ; en effet les remises - transferts financiers de l'épargne des migrants à leurs proches (familles, amis, communauté) restés dans le pays d'origine - deviennent une source importante de financement et de développement.

Selon la Banque Mondiale, en 2010, le montant de ces envois d'argent s'élevait à 239 milliards de dollars. Temporairement ralenti par la crise financière et économique de 2008, le volume des remises n'a cependant cessé de croître depuis le début des années 2000⁴. Ainsi, le rapport sur les inégalités mondiales 2018 montre que depuis 1980, les populations les plus pauvres du monde (c'est à dire les 50% du bas de la distribution des revenus) ont vu leurs revenus augmenter de manière significative, grâce à l'essor des pays émergents⁵.

Aujourd'hui il n'y a jamais eu aussi peu de personnes vivant dans la pauvreté si l'on rapporte ce chiffre à la population mondiale. Néanmoins on constate depuis plusieurs décennies que les Etats des pays les plus avancés ont du mal à éradiquer la misère au sein même de leur pays : c'est le cas des pays les plus riches en terme de PIB dans le monde (Etats-Unis, Chine, Japon...).

² <http://www.sciencespo.fr/ceri/fr/carto/dossier-carto-la-pauvrete-dans-le-monde> (consulté le 24/09/2019)

³ <https://canada.lareference.ca/en/limmigration/> (consulté le 27/09/2019)

⁴ <http://www.sciencespo.fr/ceri/fr/carto/dossier-carto-la-pauvrete-dans-le-monde> (consulté le 24/09/2019)

⁵ <http://ses.ens-lyon.fr/actualites/rapports-etudes-et-4-pages/rapport-sur-les-inegalites-mondiales-2018-wid-world-decembre-2017> (consulté le 26/09/2019)

De plus, bien que les écarts de pauvreté entre les pays développés et les pays en développement se soient considérablement réduits au cours des cinquante dernières années, ils demeurent importants et restent une préoccupation phare des dirigeants politiques. Il est donc primordial de savoir quels facteurs contribuent à réduire ou accroître la pauvreté dans un pays pour pouvoir expliquer rationnellement les écarts de niveau de pauvreté entre les pays.

Pour ce faire nous étudierons le taux de pauvreté en 2015, en appliquant l'Analyse des Composantes Principales (ACP) sur un échantillon de 59 pays à travers le monde. Avant cela, nous allons présenter notre base de données et justifier chacune des variables utilisées pour ce dossier.

Partie 2 : Analyse détaillée du sujet

A) Une variable illustrative

La variable que nous avons choisi d'expliquer dans notre analyse est le taux de pauvreté dans 59 pays du monde en 2015.

Comme l'explique l'Insee, "le taux de pauvreté correspond à la proportion d'individus dont le niveau de vie est inférieur pour une année donnée à un seuil, dénommé "seuil de pauvreté".⁶ Le niveau de vie désigne le revenu disponible, c'est à dire les revenus d'activité (nets des cotisations sociales), les revenus du patrimoine, les transferts en provenance d'autres ménages et les prestations sociales (y compris les pensions de retraite et les indemnités de chômage), nets des impôts directs.⁷

Ainsi, on dit d'un individu qu'il est pauvre s'il vit dans un ménage dont le niveau de vie est inférieur au seuil de pauvreté. Ce dernier peut être défini de manière absolue (en fonction d'un panier de consommation minimale) ou relative (en pourcentage du revenu médian ou moyen). Le choix du seuil pris en compte dans l'analyse de la pauvreté est très important, puisque cette analyse peut amener à des résultats radicalement différents selon l'option de calcul choisie.⁸

L'Insee, tout comme l'Eurostat et les autres pays européens, mesure la pauvreté monétaire de manière relative alors que d'autres pays (comme les États-Unis ou le Canada) ont une approche absolue.

Pour déterminer le seuil de pauvreté absolu, on cherche à établir le coût moyen de la consommation de base d'un adulte sur une année. Cette approche se fonde sur l'évaluation de la dépense minimale nécessaire pour assurer un niveau de vie "tolérable".

⁶ <https://www.insee.fr/fr/metadonnees/definition/c1320> (consulté le 29/09/2019)

⁷ <https://www.insee.fr/fr/metadonnees/definition/c1458> (consulté le 29/09/2019)

⁸ https://www.inegalites.fr/Les-seuils-de-pauvrete-en-France?id_theme=15 (consulté le 27/11/2019)

Dans ce dossier nous allons retenir le seuil dit “relatif”. Cette méthode de calcul tient compte du niveau de vie d'un pays : la pauvreté étudiée est fonction du niveau de revenu médian de la population.⁹ Dans les données que nous avons récoltées sur le taux de pauvreté sur les sites officiels de l'OCDE et de la Banque Mondiale, le seuil fixé s'élevait à 50% du revenu médian, cela signifie qu'est considéré comme pauvre tout individu ayant un revenu disponible inférieur à la moitié du revenu médian.

De même, la pauvreté peut être reflétée au travers de l'évolution de la distribution des revenus. Il est toutefois très difficile de définir et de mesurer la “pauvreté” en soit, en particulier lorsque l'on procède à des comparaisons internationales. En effet, les besoins de la population, ainsi que les réseaux sociaux et familiaux, ne sont pas les mêmes d'un pays à l'autre, et certains services publics gratuits dans un pays peuvent être payants ailleurs. Ainsi, un niveau de revenu donné peut se traduire par un niveau de vie différent selon les pays¹⁰. Par manque de données ou tout simplement par le fait que ces subtilités de la pauvreté sont difficilement quantifiables et comparables entre les pays, nous ne les inclurons pas dans notre analyse.

Nous essayerons cependant de donner un aperçu le plus précis possible de la pauvreté dans les 59 pays concernés, avec toutes les données dont nous avons pu disposer ; nous chercherons à déterminer les composantes principales de la pauvreté.

⁹ https://www.inegalites.fr/Les-seuils-de-pauvrete-en-France?id_theme=15 (consulté le 27/11/2019)

¹⁰ <https://www.cairn.info/revue-economique-de-l-ocde-2002-1-page-7.htm> (consulté le 07/10/2019)

B) Des variables actives

Afin d'expliquer les niveaux de taux de pauvreté dans les pays observés nous allons étudier 9 variables qui, en théorie, ont une incidence sur la pauvreté au niveau global d'un pays. Certaines de ces variables font référence à la pauvreté monétaire des individus (PIB par habitant, salaire moyen), d'autres sont davantage liées à la qualité de vie et à la santé (IDH, espérance de vie, part des dépenses de santé dans le PIB). Finalement d'autres sont liées plus indirectement à la pauvreté des individus d'un pays (indice de corruption, coefficient de Gini, taux de chômage, part de la recherche et développement dans le PIB).

Les données que nous avons récoltées proviennent essentiellement des bases de données de l'OCDE¹¹ et de la banque mondiale¹². À partir de ces informations nous avons créé notre propre base de données¹³, avec laquelle nous avons pu réaliser notre analyse (graphiques élaborés nous mêmes) et tirer nos conclusions.

Pour chaque variable active nous avons élaboré un graphique de corrélation avec la variable illustrative pour pouvoir justifier de manière visuelle le choix de nos variables :

- une corrélation positive, où la droite de régression serait croissante, signifie que lorsque la variable active étudiée augmente, la pauvreté fera de même.

- une corrélation négative, où la droite de régression serait décroissante, signifie que lorsque la variable active étudiée augmente, la pauvreté quant à elle diminuera.

De plus, nous avons trouvé pertinent de distinguer nos observations par continent pour rendre compte des disparités de la pauvreté entre ceux-ci.

Nous voulions intégrer dans nos variables explicatives la part de la population des plus de 25 ans ayant un diplôme de l'enseignement supérieur (master ou équivalent). Cette variable aurait été intéressante puisqu'elle prend en compte l'impact de l'éducation sur la pauvreté d'un pays. Depuis les années 1970, l'accès à l'éducation est une condition nécessaire pour avoir un travail avec une rémunération suffisamment élevée, et pour sortir de la pauvreté monétaire. De nos jours, cette éducation est absolument indispensable et elle ne

¹¹ <https://data.oecd.org/fr/> (consulté le 10/2019)

¹² <https://donnees.banquemondiale.org/> (consulté le 10/2019)

¹³ Notre base de données finale :

https://docs.google.com/spreadsheets/d/e/2PACX-1vRs8ljziBAAeFhiOgCImbc6iMVn88lhI4LjPz4JxAtelNLOxnAy748VRHGuu_vBft1liMID_ARo1K6i/pub?output=xlsx

garantit plus automatiquement l'accès à un travail bien rémunéré : dans les pays développés comme en développement, il est nécessaire d'avoir un diplôme de l'enseignement supérieur pour bénéficier d'une rémunération convenable. Nous n'avons cependant pas trouvé assez de données pour l'année 2015, c'est pourquoi nous n'intégrerons pas cette variable à notre analyse.

De même, un bon moyen d'appréhender le niveau de pauvreté serait de s'intéresser à deux facteurs : les écarts de revenus marchands et l'impact de la fiscalité et de la politique sociale sur la redistribution. Nous aurions aimé nous focaliser sur les effets des impôts et transferts en ajoutant des données sur le type de redistribution dans chaque pays. Ainsi nous aurions pu noter les différences considérables dans l'ampleur de cette redistribution.

Concernant la distribution du revenu, l'essentiel du revenu des ménages est constitué du revenu marchand, c'est-à-dire le revenu du travail ou le rendement du capital investi. Cependant, l'État opère des prélèvements fiscaux sur ce revenu et effectue des transferts monétaires, ce qui modifie le revenu disponible. Néanmoins les données mesurant les différents niveaux de redistribution entre les pays étaient insuffisantes, c'est pourquoi nous ne les incluons pas non plus dans notre analyse.

1- Pib par habitant (X_1) et salaire moyen (X_2)

Les variables X_1 et X_2 sont liées puisqu'il s'agit du PIB par habitant et du salaire moyen des pays concernés. Ces deux variables visent à donner un aperçu du niveau de revenu moyen dont disposent les individus des différents pays, qui par association devrait nous donner un aperçu approximatif de la pauvreté des pays.

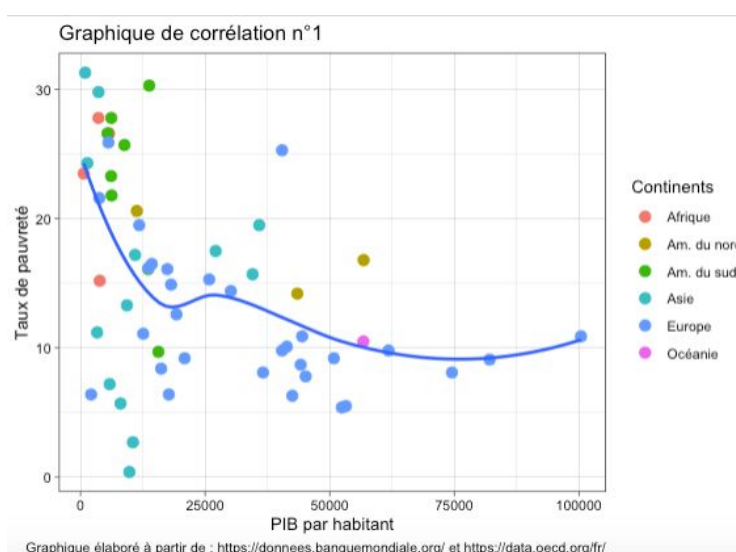
→ Le **produit intérieur brut par habitant** correspond à la valeur du PIB divisée par le nombre d'habitants d'un pays. Il est donc plus efficace que le PIB pour mesurer le développement d'un pays, cependant, il n'est qu'une moyenne donc il ne permet pas de rendre compte des inégalités de revenu et de richesse au sein d'une population.¹⁴

¹⁴ <https://www.lafinancepourtous.com/decryptages/politiques-economiques/theories-economiques/pib/>
(consulté le 08/10/2019)

Les données sur le PIB par habitant des pays pris en compte dans notre analyse sont toutes converties en dollars US de telle sorte que le pouvoir d'achat réel des différentes devises puisse être corrigé grâce à la Parité de Pouvoir d'Achat (PPA).

Selon l'office fédéral de la statistique Suisse, la croissance du PIB par habitant dépend principalement de deux facteurs. Premièrement, la productivité du travail qui représente « *le rapport entre le PIB et le nombre d'heures effectivement travaillées sur le territoire économique considéré* ». Deuxièmement, « *le nombre moyen d'heures effectives de travail par habitant* » sur une année. Suivant cette logique, ce serait donc le travail qui influencerait essentiellement le revenu par habitant.¹⁵

GRAPHIQUE N°1 : Relation entre le PIB par habitant et le taux de pauvreté



Grâce au graphique n°1 on remarque une corrélation négative assez intuitive : lorsque le PIB/habitant - c'est à dire la richesse détenue par un individu, augmente, alors le taux de pauvreté diminue dans le pays concerné. Il est intéressant de noter les différences de pauvreté et de PIB/habitant en fonction du continent → les pays d'Afrique, d'Amérique du sud et certains d'Asie ont un taux de pauvreté relativement élevé pour un PIB/habitant plutôt faible. À contrario, les pays ayant un PIB par habitant élevé et un taux de pauvreté bas sont principalement des pays d'Europe et d'Océanie (il s'agit de l'Australie).

→ La seconde variable active est le niveau de **salaire moyen** dans le pays. Il est obtenu en divisant la masse salariale totale issue des comptes nationaux par le nombre moyen de

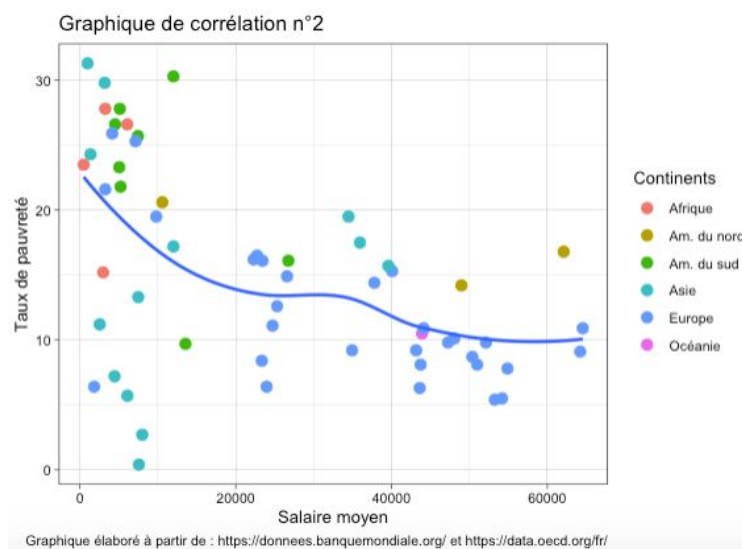
¹⁵ <https://www.ifrap.org/etat-et-collectivites/pib-par-habitant-quelle-tendance-en-europe-et-en-france>
(consulté le 09/10/2019)

salariés dans l'ensemble de l'économie, puis en multipliant le chiffre obtenu par le nombre d'heures hebdomadaires moyennes habituelles pour un salarié à temps plein rapporté à la moyenne des heures hebdomadaires travaillées pour tous les salariés. Cet indicateur est mesuré en prix constants en USD.¹⁶

Contrairement au PIB par habitant, le salaire moyen nous donne exclusivement une indication sur la rémunération du travail sans prendre en compte les autres formes de revenu dont les individus peuvent bénéficier comme les revenus fonciers et les revenus du capital.

Cet indicateur est néanmoins utile pour mesurer la pauvreté d'un pays puisqu'il permet de comparer de manière approximative les pays en terme de travailleurs pauvres. En effet, plus le salaire moyen du pays est faible, plus il serait logique qu'il y ait des travailleurs pauvres, d'autant plus que depuis les années 2000, le nombre de travailleurs pauvres augmente dans presque tous les pays développés

GRAPHIQUE N°2 : Relation entre le salaire moyen et le taux de pauvreté



Comme pour la variable « PIB/habitant », on peut noter à partir du graphique n°2 qu'il existe une corrélation négative avec notre variable illustrative. Cela valide l'hypothèse selon laquelle PIB par habitant et salaire moyen sont liés, ils donnent tous deux une approximation du niveau de vie moyen d'un pays.

2- Indice de Gini (X₃)

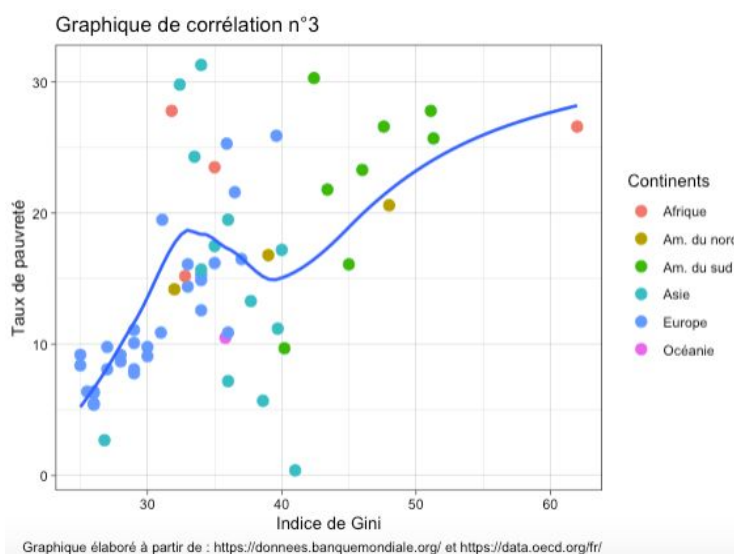
¹⁶ <https://data.oecd.org/fr/earnwage/salaires-moyens.htm> (consulté le 15/10/2019)

Les inégalités de revenu sont mesurées ici grâce au **coefficient de Gini**. Ce dernier, variant de 0 et 1, indique la répartition de la richesse entre les individus d'un territoire donné. Plus il est proche de zéro, plus la répartition des revenus est égalitaire, à contrario plus il s'approche de 1 plus la richesse est concentrée sur un petit groupe de personnes.¹⁷

Nous avons décidé d'utiliser cet indicateur comme variable explicative du taux de pauvreté parce que la pauvreté dans un pays est fortement liée aux inégalités de richesse. En effet plus la richesse est concentrée sur certains individus, moins le quartile le plus pauvre de la population aura de richesse

Il existe trois formes d'inégalités : inégalités monétaires, inégalités des conditions de vie et inégalités des chances. Dans notre analyse nous nous intéresserons exclusivement à l'inégalité monétaire. Cette dernière pose la question de la répartition des fruits de la croissance économique entre les différentes tranches de la population. Si les classes les plus riches de la population captent la majeure partie de la richesse créée, les individus les plus pauvres ne vont pas pouvoir avoir les moyens financiers de sortir de cette pauvreté même lorsque la conjoncture économique est favorable. Ici il s'agit de la structure de répartition des richesses qui influence le taux de pauvreté du pays.¹⁸

GRAPHIQUE N°3 : Relation entre l'Indice de Gini et le taux de pauvreté



D'après le graphique n°3 on remarque une relation positive assez nette entre Y et l'indice de Gini. Lorsque l'indice est proche de 0 le taux de pauvreté est lui aussi proche de 0

¹⁷ <https://data.oecd.org/fr/inequality/inegalite-de-revenu.htm> (consulté le 13/10/2019)

¹⁸ http://horizon.documentation.ird.fr/exl-doc/pleins_textes/pleins_textes_7/carton07/010025954.pdf (consulté le 13/10/2019)

puisque cela signifie que le revenu est réparti également sur le territoire ; d'où la relation positive (courbe croissante). On note que les pays développés se situent davantage en bas à gauche du graphique où les inégalités ainsi que le taux de pauvreté sont les plus faibles. En haut de la droite de régression on voit que les continents s'y trouvant sont : Afrique et Amérique du sud.

3- Indice de corruption (X4)

Depuis 1995, l'ONG Transparency International publie chaque année un **indice de perception de la corruption** (CPI) classant les pays selon le degré de corruption perçue dans un pays. L'indice est élaboré à l'aide d'enquêtes réalisées auprès d'hommes d'affaires, d'analystes de risques et d'universitaires résidant dans un pays ou à l'étranger.¹⁹ Il sera proche de 0 si le niveau de corruption est élevé, et proche de 1 s'il y a peu ou pas de corruption.

La corruption, l'inégalité et la pauvreté forment un cercle vicieux auto-entretenu par la distribution inégale du pouvoir dans la société et la répartition inégale de la richesse. De fait, la corruption perpétue la pauvreté : elle biaise les décisions et détourne des fonds déjà peu nombreux, en refusant aux populations pauvres l'accès aux services sociaux de base ainsi qu'aux ressources qui leur permettraient d'améliorer leurs conditions de vie. Si les populations pauvres ne peuvent défendre leurs droits et réclamer que leurs dirigeants rendent des comptes, leurs besoins ne sont pas remplis et ils ne peuvent échapper à la pauvreté.²⁰ Ils se retrouvent ainsi bloqués dans une spirale injuste et dangereuse.

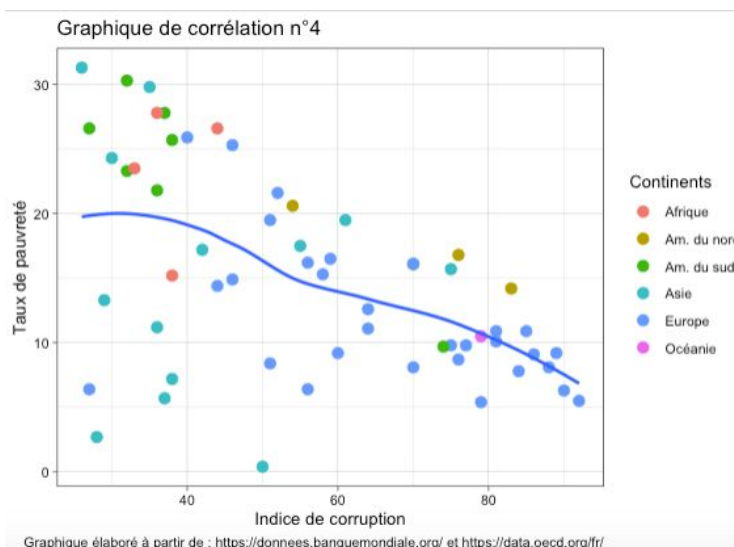
La corruption entrave le système de redistribution des richesses ; facteur important de lutte contre la pauvreté et une des fonctions principales des Etats. En effet si un Etat n'est pas en mesure d'assurer une redistribution des richesses équitable, les inégalités entre les plus fortunés et les plus pauvres risquent de se creuser. Ainsi les ménages les moins fortunés qui ne perçoivent pas ou peu de revenus liés à leur capital ne pourront plus compter que sur les revenus des parents, qui en temps de crise ou en cas de risque social (accidents de travail, vieillesse, maladie, grossesse) réduisent drastiquement les ressources financières du ménage.

¹⁹ <https://transparency-france.org/publications/indices-de-perception-de-corruption/#.Xd5Ez-TsbiU> (consulté le 16/10/2019)

²⁰ https://www.transparency.org/files/content/activity/2012_PCA_brochure_FR.pdf (consulté le 16/10/2019)

Par conséquent, la corruption accroît la vulnérabilité des ménages à la pauvreté en cas de perte de revenus, et empêche les ménages les plus pauvres d'en sortir, puisque l'Etat ne peut pas venir en aide de manière efficace à travers les transferts sociaux comme il le devrait.

GRAPHIQUE N°4 : Relation entre l'Indice de corruption et le taux de pauvreté



On constate d'après le graphique n°4 qu'il existe une relation négative entre le taux de pauvreté et l'indice de corruption, ce qui confirme notre hypothèse d'interdépendance négative entre ces deux variables.

4- Indice de développement humain (X₅)

L'**indice de développement humain** (IDH) est une mesure créée par le Programme des Nations Unies pour le Développement (PNUD). Il permet de mieux mesurer les inégalités mondiales et de trouver des solutions aux problèmes d'éducation et de pauvreté. L'indice de développement humain se mesure à partir de trois critères principaux : le produit intérieur brut (PIB) par habitant, l'espérance de vie des citoyens d'un Etat et le niveau d'éducation mesuré à partir de 15 ans et plus. En faisant entrer l'éducation et l'espérance de vie de la population dans sa grille de lecture, cet indice de mesure permet d'être plus précis dans l'analyse du développement des Etats et de leur population.

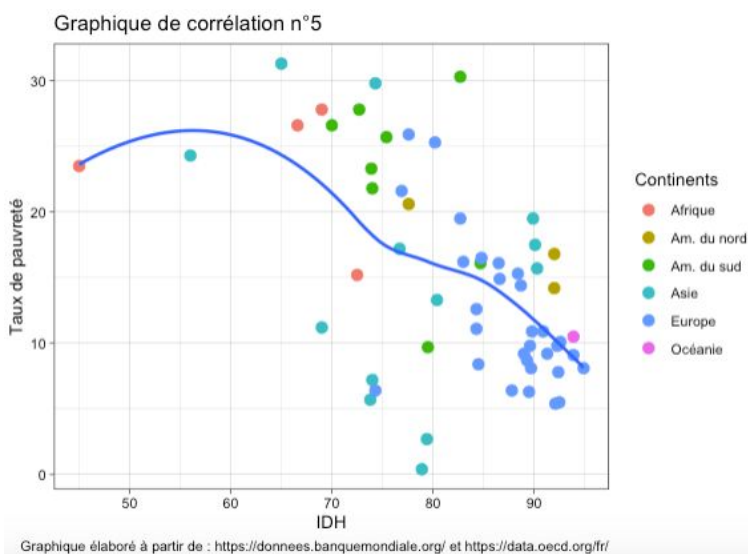
L'indice de développement humain permet d'établir chaque année un classement rendu public par le Programme des Nations Unies pour le développement. Il permet de constater des évolutions effectuées par les différents pays du monde en matière de développement.²¹

²¹ <https://www.geo.fr/environnement/idh-quest-ce-que-lindice-de-developpement-humain-193546> (consulté le

Comme expliqué ci-dessus, l'IDH a pour objectif de mesurer le niveau de développement des pays. Cependant il ne s'arrête pas au simple poids économique des pays ou à leur niveau de vie - respectivement mesurés par le PIB et le PIB/habitant - mais il intègre des données plus qualitatives telles que l'éducation par exemple.²² Cette variable est donc particulièrement intéressante puisqu'elle inclut une dimension éducative que nous n'avons pas pu intégrer dans notre analyse à cause du manque de données disponibles sur la variable "part de la population de plus de 25 ans possédant un diplôme de l'enseignement supérieur".

L'IDH se présente comme un nombre compris entre 0 et 1 : plus l'indice est proche de 1, plus le niveau de développement du pays est élevé. Bien que celui-ci ait considérablement progressé dans toutes les régions du monde entre 1990 et 2015, une personne sur trois vit toujours dans de faibles conditions de développement humain.²³

GRAPHIQUE N°5 : Relation entre l'IDH et le taux de pauvreté



D'après le graphique n°5 on remarque une corrélation négative entre Y et la variable 'IDH' : au delà de 65-70% de niveau de développement, le taux de pauvreté a tendance à être plus faible dans les 59 pays étudiés. On note aussi une certaine concentration des pays en terme d'IDH en fonction du taux de pauvreté → la grande majorité des pays se situent dans la partie droite du graphique, où l'IDH est compris entre 70 et 95%. De plus on peut remarquer qu'un pays se détache particulièrement des autres par son IDH très faible : il s'agit de l'Ethiopie dont l'indice s'élève à seulement 45%.

5- Taux de chômage (X.)

Selon le Bureau International du Travail (BIT), un chômeur est une personne en âge de travailler qui répond simultanément à trois conditions :

- être sans emploi, c'est à dire ne pas avoir travaillé au moins une heure durant une semaine de référence
- être disponible pour prendre un emploi dans les 15 jours
- avoir cherché activement un emploi dans le mois précédent ou en avoir trouvé un qui commence dans moins de trois mois.²⁴

Le **taux de chômage** correspond donc à la part de la population active étant au chômage et cherchant activement un emploi ; ils sont considérés “actifs inoccupés”.

Il est aisé de faire le lien entre la pauvreté et l'inactivité. Lorsqu'un individu est inactif il ne perçoit pas de revenus d'activité, alors que pour la majeure partie de la population, leurs ressources financières proviennent essentiellement du revenu lié au travail (surtout pour les individus les plus pauvres).

Rappelons qu'une personne est considérée pauvre si son niveau de vie après transferts et prélèvements est inférieur au seuil de pauvreté. Or la part des ménages dont aucun membre n'occupe un emploi a fortement augmentée dans la plupart des pays développés depuis les années 2000, ce qui entraîne une baisse mécanique des ressources financières dont dispose les ménage et les expose davantage à la pauvreté.

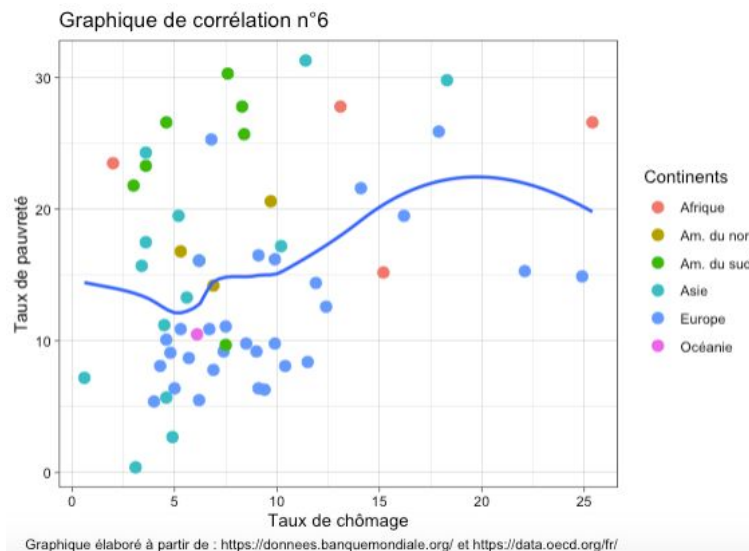
Cette évolution est tout d'abord liée à la réduction globale du temps d'activité, du fait de l'explosion du nombre de contrats à durée déterminée (CDD) et des contrats à temps partiels dans les pays du Nord, mais également à la difficulté dont font preuve ces pays à résorber le chômage de longue durée qui perdure même lorsque que la conjoncture économique est favorable.

La question serait alors de savoir comment cette exposition à des périodes d'inactivités se répercute sur le risque de pauvreté. De plus, dans la plupart des pays du monde les transferts sociaux sont insuffisants pour permettre aux individus de vivre décemment et d'échapper à la pauvreté, c'est pour cela que les actifs occupés, notamment ceux qui occupent

²⁴ <https://www.insee.fr/fr/metadonnees/definition/c1129> (consulté le 17/10/2019)

un emploi toute l'année à temps complet, apparaissent toujours beaucoup moins exposés à la pauvreté monétaire que les chômeurs.²⁵

GRAPHIQUE N°6 : Relation entre le taux de chômage et le taux de pauvreté



On voit à partir du graphique n°6, qu'il existe une corrélation positive entre le taux de chômage et celui de pauvreté. Il est intéressant de noter qu'avec une échelle presque similaire pour les 2 axes, la hausse du taux de pauvreté rapporté au taux de chômage est plus que proportionnelle : le taux de pauvreté augmente plus rapidement que le taux de chômage.

6- Dépenses de Santé (X_7) et espérance de vie à la naissance (X_8)

→ Tout d'abord, la variable active X_7 correspond à la part du PIB consacrée aux **dépenses de santé**, elle est donc exprimée en pourcentage du PIB. Indirectement elle donne une indication sur la part de la richesse que les individus allouent à se soigner. Cette variable est particulière car elle peut être interprétée de deux façons. On peut penser que si les dépenses de santé d'une population occupe une forte part dans le PIB du pays, cela signifie que les individus dépensent beaucoup d'argent pour être en meilleure santé, et donc de manière générale la population de ce pays serait en meilleure santé que les autres.

Une autre interprétation que l'on peut faire c'est de dire que si une population dépense beaucoup d'argent pour des dépenses de santé, cela signifie que la population est en mauvaise santé ou que les frais de santé dans le pays sont très élevés et donc pas accessibles pour une partie de la population, ou une combinaison des deux. La relation avec le taux de pauvreté serait alors dans ce cas positive. L'étude de cette variable nous aidera à comprendre

²⁵ http://www.onpes.gouv.fr/IMG/pdf/travaux2005-2006-1-2-2-chomage_remunerations_pauvrete.pdf
(consulté le 19/10/2019)

quelle est la bonne interprétation, néanmoins intuitivement nous pensons que que des dépenses de santé élevée est un facteur de bonne santé de la population.

→ La variable explicative X_8 qui correspond à l'**espérance de vie** à la naissance est également intéressante pour avoir un aperçu de l'état de santé de la population puisqu'un individu 'x' en bonne santé a tendance à vivre plus longtemps qu'un individu 'y' en mauvaise santé. Ainsi, une espérance de vie élevée dans un pays nous informe que sa population vit généralement en bonne santé.

⇒ Explications du lien entre la santé (dépenses en santé et espérance de vie) et la pauvreté :

Le lien qui existe entre l'état de santé de la population et le taux de pauvreté est subtile et comporte plusieurs approches. Tout d'abord lorsque la majeure partie de la population n'a pas accès aux soins les plus basiques ou refuse de dépenser une partie de leur revenu afin de remédier à des soucis de santé, cela affecte leur capacité à travailler. Les individus sont généralement moins productifs et plus fréquemment en arrêt maladie ce qui réduit leurs revenus salariaux car ils subissent une perte de revenu très importante en cas d'inactivité.

Cet effet est particulièrement grave dans les pays les moins développés : " L'une des grandes peurs des personnes démunies est que le soutien de famille tombe malade. Cela signifie que tout à coup, il n'y a plus ni nourriture ni argent. Il faut payer pour le faire soigner, ce qui aggrave encore la misère – il faudra peut-être vendre ce que l'on a et s'endetter. La spirale descendante est amorcée : la pénurie de nourriture entraîne la malnutrition et les enfants sont retirés de l'école car ils doivent aller travailler".²⁶

De même, l'état de santé des enfants peut avoir un impact sur leurs performances scolaires : de mauvaises conditions physiques liées à une santé dégradée peuvent provoquer des difficultés d'apprentissage, donc une diminution du temps d'études, ce qui les pénalisera à terme lorsqu'ils devront entrer sur le marché du travail. En effet, il apparaît qu'ils seront moins enclin à occuper un travail à hautes qualifications et seront obligés d'accepter des emplois moins qualifiés donc moins bien rémunérés ; ils seront alors plus sensibles au risque de chômage.

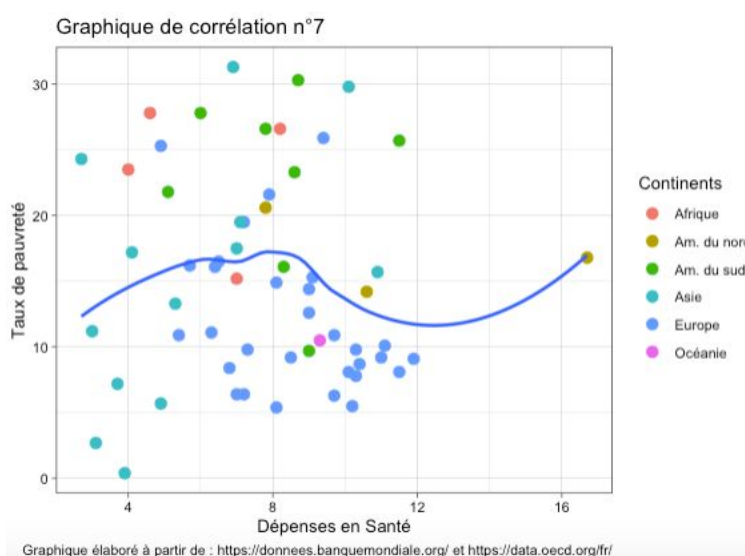
²⁶ <http://www.cafepedagogique.net/lemensuel/lenseignant/schumaines/ses/Pages/2008/Sant%C3%A9etPauvret%C3%A9danslespaysend%C3%A9veloppement.aspx> (consulté le 21/10/2019)

Nous pouvons donc conclure que les pauvres sont en partie pauvres parce qu'ils sont en mauvaise santé, on peut ajouter à cela le fait que les pauvres sont davantage exposés aux risques de santé, ce qui les place dans un cercle vicieux.

De nombreuses enquêtes le démontrent : les pauvres ont moins accès que les autres individus aux soins de base. Quand on fait la moyenne, non pondérée, des données collectées par la Banque Mondiale dans 56 pays en développement, on observe que l'accès aux soins de santé primaire augmente fortement avec le statut socio-économique du ménage.²⁷

Bien que les variables explicatives X_7 et X_8 que nous allons utiliser dans cette analyse ne mesurent pas directement le niveau de santé d'une population, elles restent pertinentes pour en donner un aperçu fiable dans beaucoup de pays.

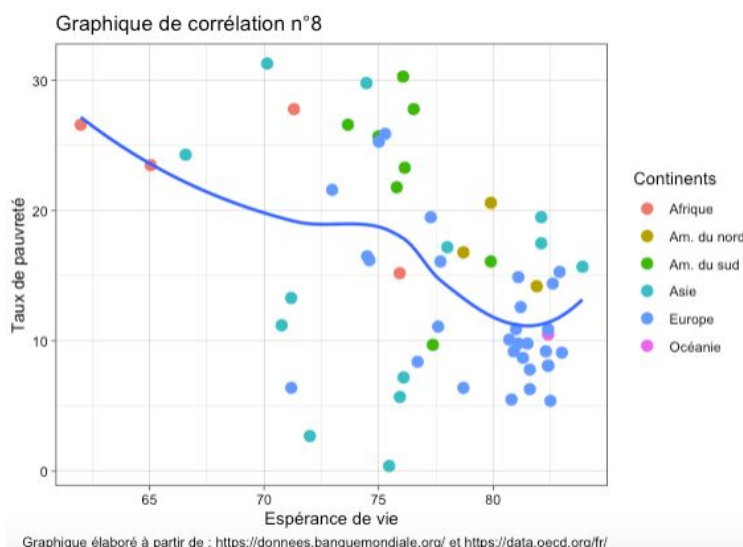
GRAPHIQUE N°7 : Relation entre les dépenses en santé et le taux de pauvreté



On voit d'après le graphique n°7 qu'il est plus difficile d'identifier la nature de la corrélation entre taux de pauvreté et dépenses de santé en % du PIB. Entre 0 et 12% du PIB dédié à la santé des individus, on retrouve des pays avec des taux de pauvreté variant entre 0 et 30%. La corrélation négative n'apparaît pas assez clairement pour valider notre hypothèse selon laquelle : lorsque les dépenses en santé augmentent, cela améliorera la santé des individus, les rendant plus aptes à travailler et à sortir de la pauvreté, formant ainsi un cercle vertueux. Les pays d'Europe semblent former un bloc où les dépenses en santé sont comprises entre 5 et 12%. A côté de cela on remarque que les Etats Unis ont le pourcentage le plus élevé de notre échantillon avec un taux à 16.7%.

²⁷ <http://www.cafepedagogique.net/lemensuel/lenseignant/schumaines/ses/Pages/2008/Sant%C3%A9etPauvret%C3%A9danslespaysend%C3%A9veloppement.aspx> (consulté le 22/10/2019)

GRAPHIQUE N°8 : Relation entre l'espérance de vie et le taux de pauvreté



D'après le graphique n°8 on remarque une relation négative entre le taux de pauvreté et l'espérance de vie à la naissance, ce qui confirme nos intuitions sur la corrélation existante entre ces 2 variables. On constate les mêmes disparités entre les continents : l'espérance de vie est élevée (+ 80 ans) pour les pays d'Europe principalement, d'Océanie et d'Asie. Alors qu'elle est plutôt faible (- 75 ans) pour les pays d'Afrique, d'autres d'Asie et d'Amérique du Sud. Ici encore nous pouvons identifier les 3 pays ayant une espérance de vie particulièrement faible (inférieure à 70 ans) : Afrique du Sud, l'Ethiopie et le Pakistan.

7- Dépenses intérieures brutes de recherche et développement (X₉)

Les dépenses intérieures de recherche et développement, aussi appelées **DIRD** correspondent, selon l'INSEE "aux travaux de recherche et développement (R&D) exécutés sur le territoire national quelle que soit l'origine des fonds. Une partie est exécutée par les administrations, l'autre par les entreprises. Elle comprend les dépenses courantes (masse salariale des personnels de R&D et dépenses de fonctionnement) et les dépenses en capital (achats d'équipements nécessaires à la réalisation des travaux internes à la R&D et opérations immobilières réalisées dans l'année).²⁸

Cet indicateur regroupe donc les dépenses publiques comme privées et est exprimé en pourcentage du PIB. Il existe trois types d'activités au sein des DIRD : la recherche fondamentale (travaux de recherche théoriques), la recherche appliquée (discerner les

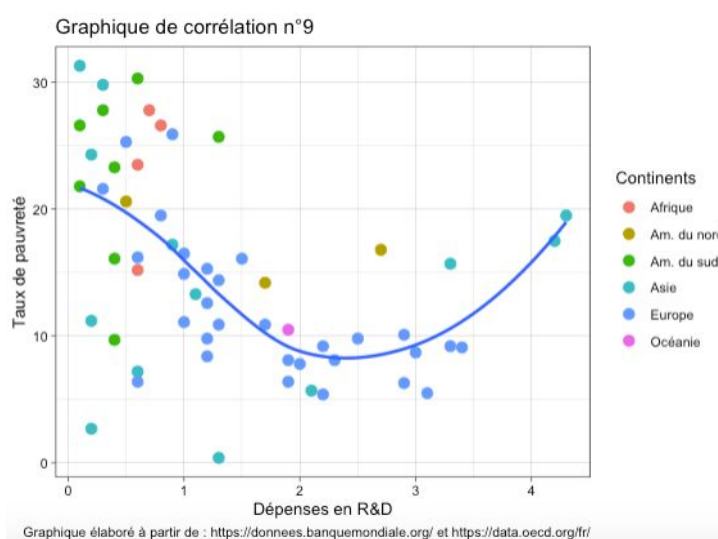
²⁸ <https://www.insee.fr/fr/metadonnees/definition/c1226> (consulté le 28/10/2019)

applications possibles du type n°1) et le développement expérimental (travaux systématiques réalisés pour lancer la fabrication).²⁹

Les travaux de recherche et développement financés par les dépenses en R&D peuvent avoir de fortes répercussions sur le niveau de vie d'un pays : ils permettent d'accroître les connaissances dans de nombreux domaines d'application tels que la science, les technologies mais aussi la culture, l'humain et la société.

De telles connaissances fournissent une instruction générale, elles renforcent les capacités et les compétences des individus et permettent donc de sortir de l'ignorance, souvent facteur de pauvreté. Ce mécanisme peut, pour nous, être résumé par ce vieux dicton anglais : "Knowledge is power", la connaissance donne le pouvoir, celui de sortir de la pauvreté. Prenons un exemple de travaux financés par le gouvernement sur la lutte contre le cancer ; cela entre bien dans les dépenses de R&D. Imaginons que, à terme, les travaux conduisent à la trouvaille d'un vaccin permettant d'éradiquer le cancer, alors, les dépenses initialement placées en laboratoires auront permis d'améliorer la santé des habitants du pays, réduisant ainsi la précarité grâce à l'amélioration des conditions de vie. Finalement la relation entre dépenses en recherche et développement et taux de pauvreté apparaît plus clairement : nous pensons qu'il s'agit d'une relation négative.

GRAPHIQUE N°9 : Relation entre les dépenses R&D et le taux de pauvreté



Finalement, à partir du graphique de corrélation n°9 nous pouvons vérifier l'hypothèse émise précédemment : il existe bien une relation négative entre les dépenses intérieures

²⁹ <https://www.enseignementsup-recherche.gouv.fr/reperes/telechar/res/res00/rapmeto00.pdf> (consulté le 28/10/2019)

brutes de R&D et le taux de pauvreté. Cependant nous observons que la droite de régression a tendance à “se redresser” à la fin - cela est dû aux 2 pays d’Asie qui semblent être atypiques.

⇒ Pour terminer cette partie explicative de nos variables nous allons présenter un tableau récapitulatif reprenant, pour chaque continent, les caractéristiques visuellement observables pour chacune des variables selon nos graphiques de corrélation.

TABLEAU N°1 : Récapitulatif des corrélations entre les différents X avec Y

	Afrique (4 pays)	Am.Nord (3 pays)	Am.Sud (8 pays)	Asie (13 pays)	Europe (30 pays)	Océanie* (1 pays)
PIB/hab	très faible	faible et modérée	plutôt faible	très faible et modérée	répartition homogène (faible à élevée)	élevée
Salaire moyen	très faible	faible et élevée	très faible et modérée	très faible et modérée	modérée et élevée	élevée
Indice de Gini	modérée et très élevée	modérée	élevée	faible et modérée	modérée	plutôt faible
Indice de corruption	faible	modérée et élevée	faible	faible et modérée	modérée et élevée	élevée
IDH	très faible et faible	modérée et élevée	faible et modérée	faible, modérée, élevée	modérée et élevée	très élevée
Taux de chômage	faible, modérée et élevée	faible	faible	faible et modérée	faible, modérée et élevée	faible
Espérance de vie	très faible et modérée	élevée	modérée	très faible, faible et élevée	modérée et élevée	élevée
Dépenses Santé	faible et modérée	modérée et très élevée	modérée	faibles, modérée et élevée	modérée et élevée	modérée
Dépenses R&D	faible	faible et modérée	faible	faible et modérée	faible, modérée et élevée	modérée

* Notre base de données comporte seulement 1 pays faisant partie de l'Océanie : l'Australie. Bien que celui-ci représente 90% du continent total en terme de superficie, l'échantillon reste peu représentatif donc nos conclusions seront très approximatives.

Au vu des résultats du tableau nous pouvons classer les continents en 4 principaux groupes :

- Pauvre/défavorisé : Afrique
- En voie de développement : Amérique du Sud
- Aisés : Asie, Amérique du Nord
- Riches/développés : Europe et Océanie

On remarque donc de fortes disparités au sein de notre échantillon. De plus il semble y avoir une grande hétérogénéité dans les pays d'Asie comme d'Europe ; pour la plupart des variables on retrouve des valeurs étant faibles, moyennes, ou élevées au sein d'une même variable.

A contrario on observe une plus grande homogénéité en Amérique du Nord comme du Sud, et en Afrique (Océanie exclu des analyses d'homogénéité car 1 seul pays représenté). Cette tendance semble se vérifier pour toutes les variables excepté le taux de chômage ; effectivement les caractéristiques observées précédemment pour chaque continent semblent rompues : les plus forts taux de chômage se trouvent en Europe (ainsi qu'un pays d'Afrique), alors que les plus faibles proviennent d'Afrique, d'Asie et d'Amérique du Sud. Cela semble donc "casser" la tendance selon laquelle nous avons classé nos pays en 4 catégories.

Nous pouvons supposer que cela est dû à l'intégration économique des pays dans la mondialisation. En effet nous retrouvons des taux de chômage faibles pour les pays des continents classés selon notre analyse en tant que "pauvres" ou "en voie de développement". Ceux-ci ont été moins impactés par les retombées économiques de la crise de 2008 initialement américaine (faillite de Lehman Brothers) qui s'est répandue rapidement à travers le monde. Il y eu des conséquences internationales liées à la globalisation qui crée une sorte d'interdépendance entre les pays du monde. L'Afrique, l'Amérique du Sud et l'Asie étant légèrement moins intégrés dans cette mondialisation, ont moins subi la récession économique qui a suivi la crise ; c'est pourquoi leur taux de chômage en 2015 est plus faible que dans les autres pays dits plus "intégrés". *Cela reste une hypothèse émise à partir de nos connaissances et des observations de notre analyse.*

Partie 3 : Analyse en Composantes Principales (ACP)

Dans l'ACP les variables déterminent les axes ; elles se présentent sous forme de flèches dans un cercle unitaire de corrélations, plus elles sont importantes dans l'explication d'un axe, plus elles seront positionnées proches de cet axe et aux extrémités.

Si le nombre d'axes "p" est supérieur à 3, il sera très difficile de faire une présentation graphique des composantes, c'est pourquoi nous retiendrons seulement deux à trois axes en fonction de l'inertie expliquée par chacun d'entre eux. Le but de cette analyse est donc de faire ressortir les variables les plus importantes afin de définir les variables latentes, aussi appelées 'composantes principales'. Ces dernières représentent une combinaison linéaire des variables explicatives initiales, nous aurons donc 9 dimensions avec une inertie totale égale à 100%. Nous essayerons de perdre le moins d'information possible en cherchant une inertie élevée tout en restreignant le nombre de composantes pour les définir au mieux.

Dans un second temps nous nous intéresserons au positionnement des individus sur les axes, qui doit confirmer l'existence des variables latentes. Nous chercherons les éléments originaux car étant "extrêmes", ils nous aideront d'avantage à confirmer les variables latentes.

A) Valeurs propres et nombre d'axes

1- Analyse descriptive

➤ Pour commencer il convient de faire quelques statistiques descriptives de base sur chaque variable. Nous étudions les variables actives mais également la variable illustrative. Cela permet d'étudier les caractéristiques de chaque variable indépendamment les unes des autres. La fonction "summary" nous donne accès aux valeurs minimales et maximales de chaque variable, la moyenne, la médiane et les quartiles.

TABLEAU N°2 : Statistiques descriptives de nos variables

<code>> summary(basepauvreté[,3:12])</code>									
Tx_pauv		PIB_hab		Sal_moy		Gini		Corrup	
Min.	: 0.40	Min.	: 639.3	Min.	: 478.7	Min.	: 25.00	Min.	: 26.00
1st Qu.:	9.15	1st Qu.:	6201.7	1st Qu.:	5643.2	1st Qu.:	29.00	1st Qu.:	37.50
Median	:14.40	Median	: 15613.8	Median	:23344.0	Median	:34.00	Median	:55.00
Mean	:15.12	Mean	: 24728.2	Mean	:24873.5	Mean	:34.98	Mean	:56.41
3rd Qu.:	21.10	3rd Qu.:	40917.5	3rd Qu.:	43674.5	3rd Qu.:	38.80	3rd Qu.:	75.50
Max.	:31.30	Max.	:100428.0	Max.	:64551.5	Max.	:62.00	Max.	:92.00

IDH	Tx_chom	Sante	Espce_vie	RetD
Min. :45.00	Min. : 0.600	Min. : 2.700	Min. :62.00	Min. :0.100
1st Qu.:74.85	1st Qu.: 4.850	1st Qu.: 6.150	1st Qu.:75.00	1st Qu.:0.600
Median :84.30	Median : 6.900	Median : 7.900	Median :77.70	Median :1.200
Mean :82.03	Mean : 8.468	Mean : 7.829	Mean :77.48	Mean :1.403
3rd Qu.:89.85	3rd Qu.:10.050	3rd Qu.: 9.700	3rd Qu.:81.55	3rd Qu.:2.050
Max. :94.90	Max. :25.400	Max. :16.700	Max. :83.90	Max. :4.300

Nous remarquons d'emblée qu'il existe de fortes disparités entre les pays, l'étendue (écart entre les valeurs minimales et maximales pour une variable donnée) est assez importante, notamment pour les variables PIB par habitant, salaire moyen ou encore taux de chômage. Cet indice étant très sensible aux valeurs extrêmes, nous supposons l'existence de valeurs atypiques pour ces 3 variables, ce que nous pourrions confirmer ou non par la suite grâce aux diagrammes de Tukey.

Nous voyons aussi que pour la grande majorité des variables, la moyenne et la médiane sont très proches. La médiane se différencie de la simple moyenne arithmétique car elle suppose que 50% de l'échantillon a des valeurs inférieures à celle-ci, et 50% a des valeurs supérieures. Ainsi, lorsqu'elle est proche de la moyenne - comme c'est le cas ici - cela signifie que les données sont relativement uniformes, il n'y a apparemment pas de valeurs qui faussent la moyenne. Il apparaît donc que nos observations sont relativement homogènes même si l'étendue reste généralement élevée.

Nous pouvons aussi nous arrêter sur la variable "PIB par habitant" qui est la seule ayant une médiane éloignée de la moyenne. 50% des pays ont un PIB par habitant inférieur à 15.613,8\$ alors que la moyenne s'élève à 24.728,2\$, cela signifie qu'un ou plusieurs pays ont un PIB par habitant très fort qui a tendance à "tirer" la moyenne vers le haut. C'est dans ce genre de cas qu'il est intéressant de comparer moyenne et médiane, dont la différence a été expliquée précédemment.

➤ Nous allons maintenant vérifier l'homogénéité des variables en comparant la moyenne et l'écart type pour chacune d'entre elles. Si l'écart type est plus élevé que la moyenne alors nous parlons d'hétérogénéité, dans le cas contraire il s'agit d'homogénéité.

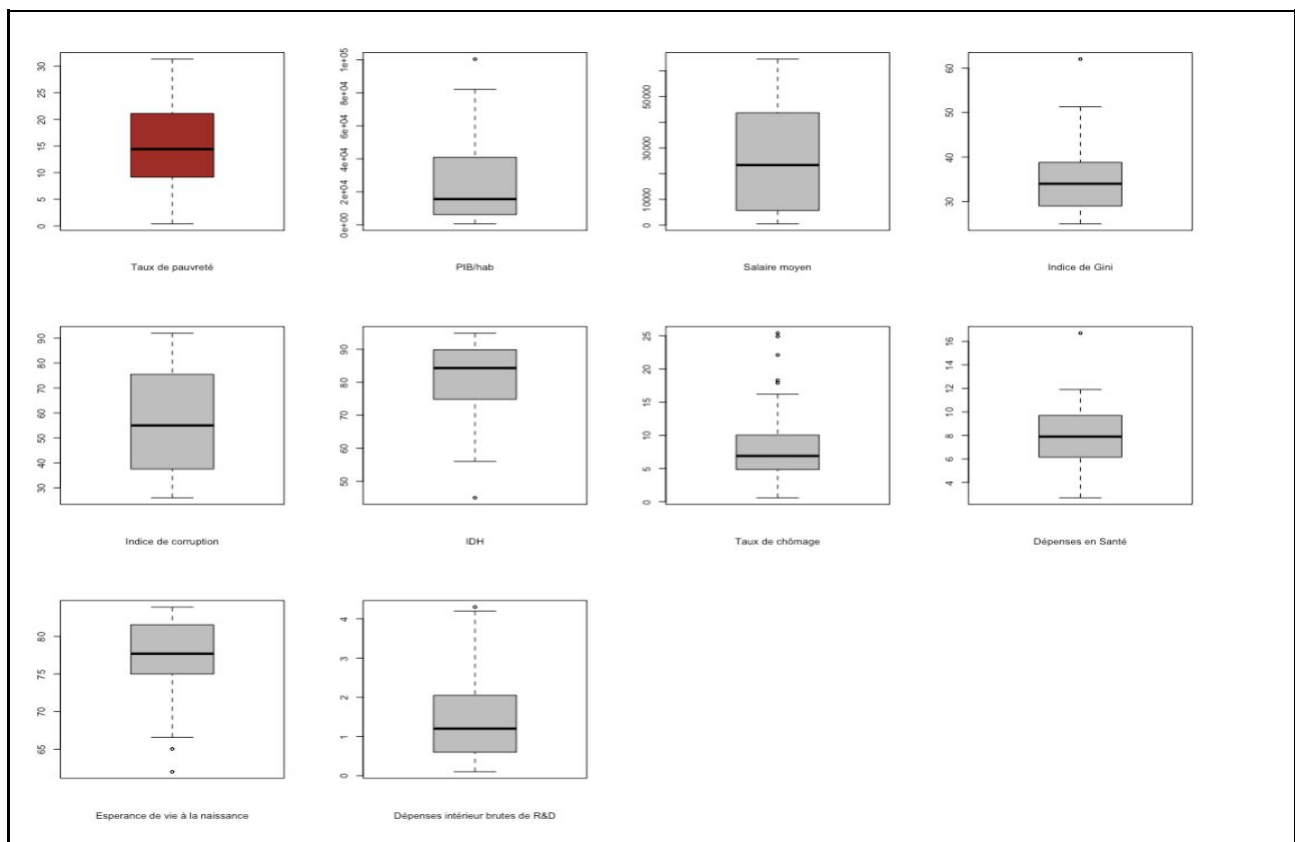
TABLEAU N°3 : Comparaison entre la moyenne et l'écart-type

<pre>> round(sapply(basepauvreté[,3:12],mean),2)</pre>										
Tx_pauv	PIB_hab	Sal_moy	Gini	Corrup	IDH	Tx_chom	Sante	Espce_vie	RetD	
15.12	24728.20	24873.45	34.98	56.41	82.03	8.47	7.83	77.48	1.40	
<pre>> round(sapply(basepauvreté[,3:12],sd),2)</pre>										
Tx_pauv	PIB_hab	Sal_moy	Gini	Corrup	IDH	Tx_chom	Sante	Espce_vie	RetD	
7.78	22680.57	20200.41	7.53	20.35	10.05	5.29	2.67	4.84	1.08	

Nous constatons que toutes nos variables sont homogènes puisque les écarts-types sont inférieurs aux moyennes. Il est donc nécessaire de centrer et réduire les variables afin de procéder à l'analyse. Lorsque nous procédons à l'analyse en composantes principales sur le logiciel R, celui-ci centre et réduit automatiquement les variables, aucune manipulation additionnelle n'est donc utile.

➤ Nous pouvons à présent regarder les diagrammes à moustaches de nos 10 variables afin de vérifier ou non, l'existence de variables atypiques.

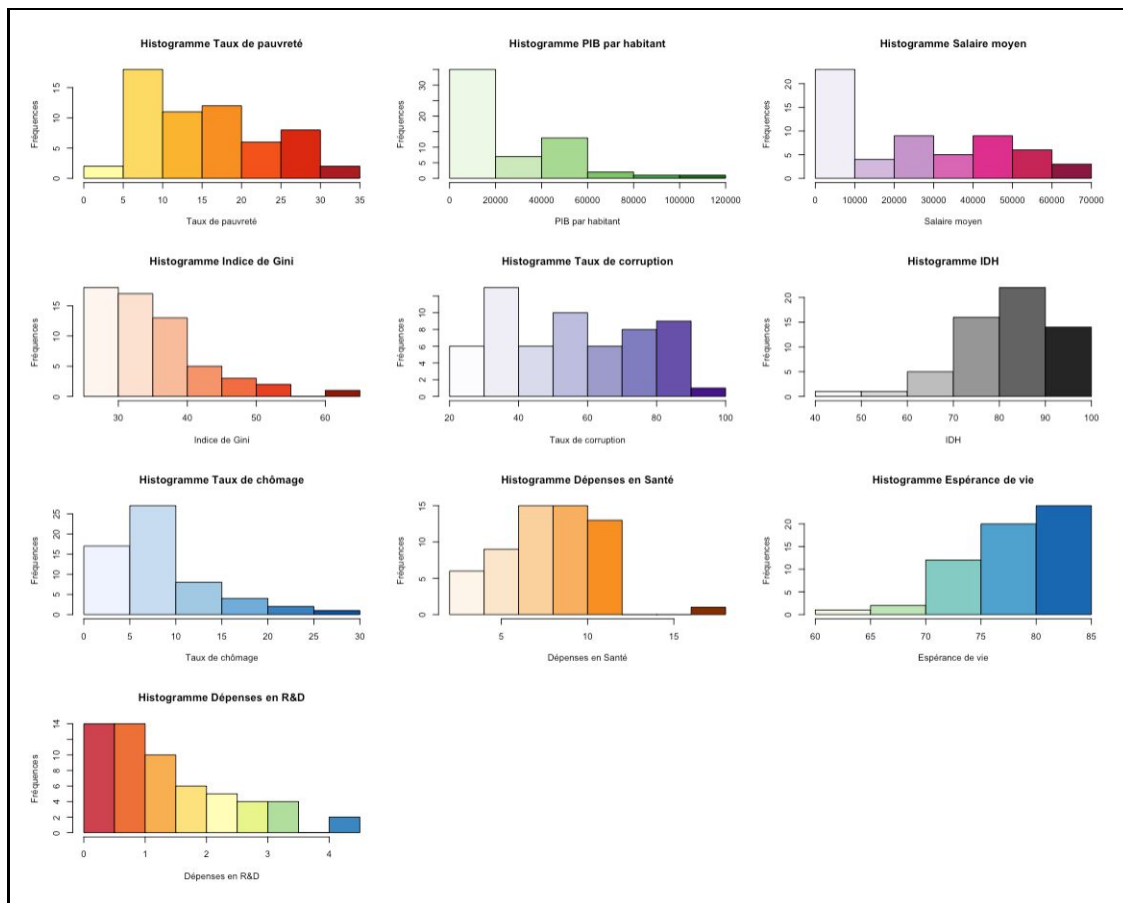
FIGURE N°1 : Diagrammes de Tukey sur nos 10 variables



Nous confirmons ainsi l'existence de valeurs atypiques, la variable taux de chômage est celle qui contient le plus de valeurs atypiques : il y en a 5. Pour les variables telles que PIB par habitant, indice de Gini, IDH, dépenses en Santé et dépenses en R&D il y en a une seule. Enfin, on retrouve 2 pays atypiques dans la variable "espérance de vie". Il est intéressant de noter que le pays atypique en terme de PIB par habitant est celui du Luxembourg qui s'élève à 100.428\$ par an et par personne. Il est en effet atypique car trop élevé par rapport au PIB/habitant des autres pays.

➤ Regardons maintenant les histogrammes des variables, ceux-ci nous donnerons une idée de la distribution de ces dernières, nous pourrons ainsi voir si leur distribution s'apparente à une loi normale. D'après la figure n°2, nous remarquons que les variables ne semblent pas suivre une loi normale. Néanmoins celles qui s'en approchent le plus sont l'IDH et le taux de chômage avec une asymétrie à droite pour IDH et une asymétrie à gauche pour taux de chômage.

FIGURE N°2 : Histogrammes des variables actives



➤ Nous pouvons compléter cette analyse en cherchant la **skewness** et la **kurtosis** des variables : ces coefficients donnent une indication sur la symétrie et l'aplatissement de la distribution des variables. La loi normale se caractérise par un coefficient d'aplatissement et d'asymétrie égal à zéro. S'ils sont différents de zéro il y a 2 situations possibles pour chaque coefficient :

- skewness positif : asymétrie à gauche, et skewness négatif : asymétrie à droite
- kurtosis positif : la distribution sera moins aplatie que la dist normale théorique, et s'il est négatif : la distribution sera plus aplatie

Le tableau n°4 ci-dessous nous confirme donc l'asymétrie des variables, plus particulièrement pour PIB par habitant, Gini, taux de chômage et IDH qui ont un skewness élevé par rapport à

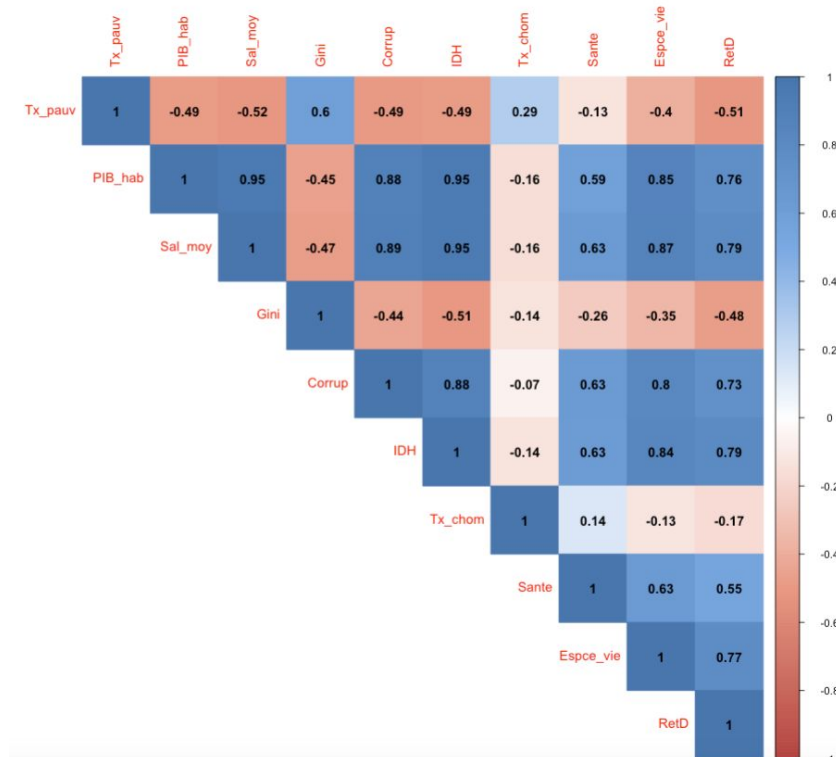
celui des autres variables. Pour les 3 premières variables citées ci-dessus l'asymétrie est à gauche, en revanche elle est à droite pour l'IDH, ce qui confirme nos hypothèses émises au regard des histogrammes. En ce qui concerne le coefficient d'aplatissement, nous constatons que pour toutes les variables leur distribution est moins aplatie que la distribution normale. Ainsi, les variables s'apparentant le plus à une loi normale sont le salaire moyen et le taux de corruption, pour lesquels les deux coefficients sont les plus proches de zéro.

TABLEAU N°4 : Coefficients d'asymétrie et d'aplatissement de chaque variable active

<code>> round(sapply(basepauvreté[,4:12],skewness),2)</code>									
PIB_hab	Sal_moy	Gini	Corrup	IDH	Tx_chom	Sante	Espce_vie	RetD	
1.19	0.38	1.12	0.16	-1.19	1.47	0.31	-1.00	0.88	
<code>> round(sapply(basepauvreté[,4:12],kurtosis),2)</code>									
PIB_hab	Sal_moy	Gini	Corrup	IDH	Tx_chom	Sante	Espce_vie	RetD	
3.98	1.72	4.60	1.67	4.81	5.08	3.62	3.78	2.94	

➤ Nous allons maintenant analyser la matrice de corrélation en regardant les coefficients ; s'ils sont compris entre 0 et 0,5 nous considérons que les variables sont faiblement corrélées, entre 0,5 et 0,6 nous pouvons supposer l'existence d'une corrélation et au delà de 0,6 nous estimons que la corrélation est forte.

FIGURE N°3 : Matrice de corrélation



→ Relation au sein des variables actives

Au regard de la figure n°3 nous remarquons que la plupart des variables sont fortement corrélées entre elles. Pour les variables PIB par habitant et salaire moyen par exemple la corrélation s'élève à 95%, ce qui est considérable.

De manière générale, les variables PIB par habitant, salaire moyen, taux de corruption, IDH, espérance vie et dépenses en R&D sont fortement corrélées. Cela devrait être confirmé visuellement sur le cercle de corrélations. En effet les flèches correspondantes à ces variables devraient toutes être proches les unes des autres et dirigées dans le même sens.

En revanche pour l'indice de Gini, les dépenses en santé ainsi que le taux de chômage, on remarque qu'elles ont une corrélation généralement plus faible avec le reste des variables. On peut alors supposer qu'elles seront toutes les trois éloignées du reste des variables, sans pour autant être proches entre elles, spécialement la variable taux de chômage qui au vu de ces chiffres de corrélation, apparaît comme indépendante des autres variables explicatives.

Nous allons comparer cette analyse avec celle de l'analyse en composante principale et plus particulièrement l'analyse du cercle de corrélation pour voir si elle confirme ou non les constats faits à partir de la matrice.

→ Relation entre les variables actives et la variable illustrative

Le tableau n°5 ci-dessous récapitule les relations entre les variables actives et la variable illustrative observables à l'aide des graphiques de corrélation et la matrice de corrélation.

TABLEAU N°5 : Récapitulatif des corrélations existantes entre les variables

	Graphiques³⁰	Matrice de corrélation³¹	Conclusion
PIB par habitant	clairement négative	négative (CC* \approx -0,5)	La matrice de corrélation confirme la trajectoire de la courbe sur le graphique.
Salaire moyen	clairement négative	négative (CC \approx -0,5)	La matrice de corrélation confirme la trajectoire de la courbe sur le graphique.

³⁰ La colonne "**Graphique**" fait référence aux graphiques de corrélation disponibles en partie 2 pour chaque variable avec le taux de pauvreté Y.

³¹ La colonne "**Matrice de corrélation**" fait référence à la matrice de corrélation disponible en figure n°3

Indice de Gini	pas clairement positive	fortement positive (CC=0,6)	La matrice de corrélation nous révèle que la corrélation est fortement positive entre les deux variables, ce qui n'était pas identifiable avec certitude sur le graphique, étant donné que la courbe devenait légèrement décroissante à un certain point.
Indice de corruption	clairement négative	négative (CC≈ -0,5)	La matrice de corrélation confirme la trajectoire de la courbe sur le graphique.
IDH	clairement négative	négative (CC≈ -0,5)	La matrice de corrélation confirme la trajectoire de la courbe sur le graphique.
Taux de chômage	intuitivement positive mais difficilement identifiable	faiblement positive (CC = 0,29)	Du fait que les deux extrémités de la courbe soient décroissantes, nous n'étions pas certain de la corrélation. Cependant la matrice de corrélation confirme l'intuition que nous avons.
Espérance de vie	clairement négative	faiblement négative (CC = -0,4)	La matrice de corrélation nous indique un CC égal à -0,4 ce qui est assez faible étant donné que la courbe était parfaitement décroissante (sauf à la fin où elle se redressait légèrement)
Dépenses Santé	non identifiable	très faiblement négative (CC = -0,13)	La matrice de corrélation nous révèle la corrélation négative entre les deux variables, qui n'était pas identifiable avec certitude sur le graphique
Dépenses R&D	intuitivement négative	négative (CC≈ -0,5)	Du fait que la droite de régression se redresse à la fin, nous n'étions pas certains de la corrélation. La matrice de corrélation confirme l'intuition que nous avons.

* : coefficient de corrélation (CC)

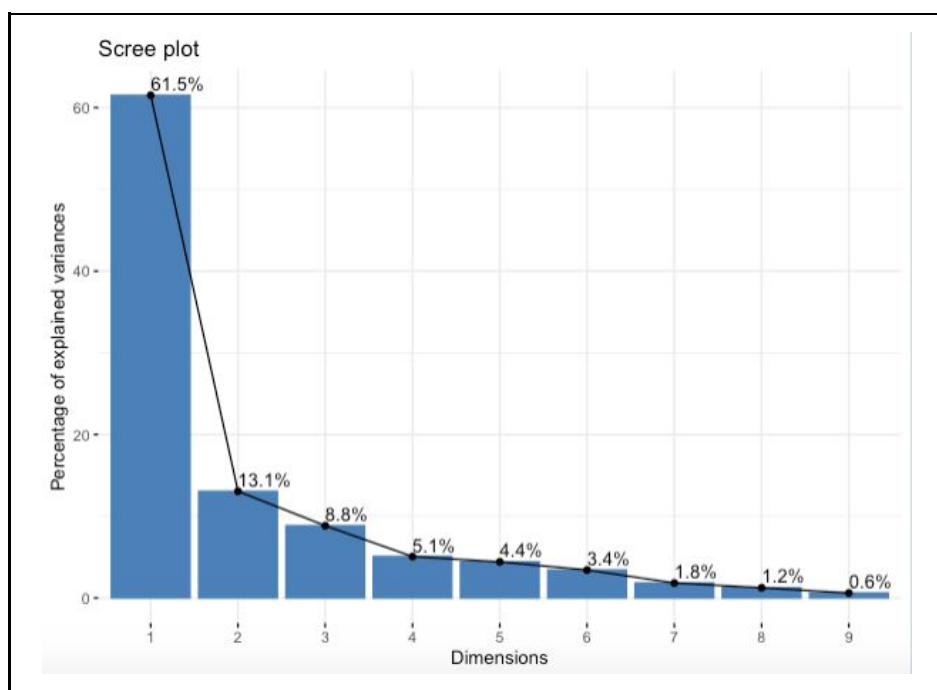
2- Valeurs propres et tableau d'inertie

TABLEAU N°6 : Valeurs propres et inertie des 9 composantes

<code>> round(res.pca\$eig,2)</code>			
	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	5.54	61.50	61.50
comp 2	1.18	13.07	74.57
comp 3	0.80	8.85	83.42
comp 4	0.46	5.07	88.49
comp 5	0.40	4.41	92.90
comp 6	0.31	3.42	96.32
comp 7	0.16	1.83	98.15
comp 8	0.11	1.24	99.39
comp 9	0.05	0.61	100.00

À partir du tableau n°6, on voit que la première composante contribue à hauteur de 61,5 % de l'inertie totale du nuage initial, et la seconde à 13,07%. Cela signifie que ces deux composantes principales permettent de conserver 74,57% de l'information totale du nuage initial si l'on décide d'effectuer un plan à deux dimensions. La question est de savoir s'il est nécessaire dans notre analyse de prendre également en compte la troisième composante principale qui explique 8,85% de l'inertie totale. Pour déterminer le nombre de composantes principales on peut regarder le graphique des valeurs propres (appelé "scree plot"). Le nombre d'axes peut être déterminé par le point, en dessous duquel les valeurs propres restantes sont toutes relativement petites et de tailles comparables.

GRAPHIQUE N°10 : Inertie associée à chaque axe



À partir du graphique n°10 ci-dessus, nous pourrions vouloir nous arrêter à la troisième composante principale. Selon le tableau n°6, 83,42% des informations (variances) contenues dans les données sont conservées par ces trois premières composantes. Cependant, l'inertie moyenne est de 1,001 (on l'obtient en faisant le rapport de l'inertie totale 9,01 par le nombre de variables).

De manière générale il est intéressant de retenir les axes dont l'inertie est supérieure à l'inertie moyenne soit 1,001, ce n'est pas le cas du troisième axe dont l'inertie est de 0,8. De plus la troisième composante représente moins de 10% de l'inertie totale (8,85%), il ne

semble donc pas nécessaire de la retenir pour tester la projection des variables et des individus sur un plan 1:3. Dans notre analyse, les deux premières composantes principales expliquent 74,57% de la variation - c'est un pourcentage acceptable.

B) Variables ; contributions, corrélations, cosinus carrés

Dans cette partie nous allons étudier la relation entre les composantes principales et nos 9 variables initiales. Il est à noter que le logiciel R donne les résultats pour les cinq premières dimensions, cependant seules les deux premières vont être analysées comme cela a été décidé dans la partie précédente.

1- Contributions

Tout d'abord nous allons étudier la contribution de chaque variable à la formation des axes 1 et 2 à l'aide de l'annexe n°1 et du tableau n°7. Celui-ci se lit en pourcentage dont la somme de chaque colonne est égale à 100.

TABLEAU N°7 : Contribution de chaque variable à la formation des axes

```
> round(res.pca$var$contrib,2)
```

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
PIB_hab	14.02	0.62	0.40	17.31	17.54
Sal_moy	16.73	0.01	0.02	1.71	5.32
Gini	5.16	10.89	64.63	8.12	2.31
Corrup	15.04	0.02	0.35	4.66	6.61
IDH	14.97	1.05	0.60	2.55	18.18
Tx_chom	0.63	63.12	23.04	0.42	0.31
Sante	8.05	23.07	8.92	18.67	4.33
Espce_vie	13.71	0.21	0.09	2.95	44.77
RetD	11.69	1.01	1.96	43.61	0.64

A partir du tableau n°7, nous allons analyser les composantes de la première dimension, puis nous passerons à l'analyse de la deuxième dimension :

➤ Dimension 1 : En observant la colonne liée à la dimension 1, nous pouvons diviser les variables en trois groupes. Le premier groupe ne contribue quasiment pas à la définition de l'axe 1 (seulement 0,63%), il est composé de la variable 'taux de chômage'. Dans un deuxième groupe nous retrouvons les variables 'indice de Gini' et 'dépenses de santé' qui ont un pourcentage de contribution relativement faible (respectivement 5,16 % et 8,05%). Enfin

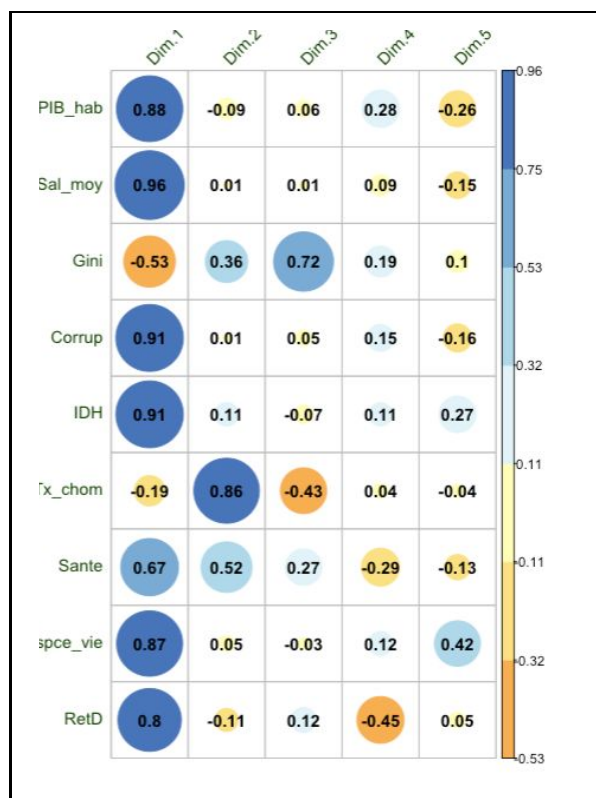
le dernier groupe est composé du reste des variables explicatives qui contribuent à plus de 10% de la construction de l'axe.

➤ Dimension 2 : Pour l'axe 2 les résultats sont totalement différents. La variable 'taux de chômage' concentre à elle seule presque les deux tiers de la dimension (63,12%), suivit par 'dépenses de santé' et 'indice de Gini' qui représentent le tiers restant (respectivement 23,07% et 10,89%). Les autres variables ne contribuent quasiment pas à la contribution de l'axe n°2 (moins de 3%).

2- Corrélations

Après l'étude des contributions nous allons étudier les corrélations entre les variables et les axes à partir du tableau n°8. Les corrélations nous indiquent la position de chaque variable initiale par rapport aux facteurs, ces corrélations vont nous aider à donner une interprétation des axes. Le signe de la corrélation est donc un élément tout aussi important que leur grandeur. Pour ce qui est des valeurs proches de 0, elles ne seront pas pertinentes quant à l'axe étudié.

TABLEAU N°8 : Corrélation de chaque variable aux 5 premières dimensions



➤ Dimension 1 : Les variables 'PIB/habitant', 'Salaire moyen', 'indice de Corruption', 'IDH', 'dépenses de santé', 'espérance de vie', et 'dépenses en R&D' sont fortement et

positivement corrélées avec l'axe 1. Les flèches représentant les variables seront orientées vers la droite le long de l'axe n°1.

➤ Dimension 2 : Les variables 'taux de chômage', et dans une moindre mesure 'dépenses de santé' et 'indice de Gini' sont corrélées positivement avec le facteur 2.

3- Cosinus carrés

Nous pouvons maintenant nous demander si les variables sont bien représentées par chaque axe. Pour ce faire, nous allons étudier le tableau n°9 des cosinus carrés des variables ainsi que les annexes n°2 et 3. Ces derniers mesurent la qualité de projection, elle-même mesurée par l'angle d'une variable avec l'axe (1 ou 2 dans notre cas) ; il nous informe donc de la présentation des variables sur le plan. Un cosinus carré élevé (proche de 1) indique une bonne représentation de la variable sur les axes principaux en considération. A l'inverse un faible cosinus carré (proche de zéro) indique que la variable n'est pas bien représentée par les axes principaux.

Pour une variable donnée, la somme des cosinus carrés sur toutes les composantes principales est égale à 1.

TABLEAU N°9 : Cosinus carrés de chaque variable aux 5 premières dimensions



Après avoir analysé le tableau n°9 ci-dessus nous pouvons souligner 3 points :

- Les variables 'PIB/hab', 'salaire moyen', 'indice de corruption', 'IDH', 'espérance de vie', et 'dépenses en R&D' sont particulièrement bien représentées par l'axe 1.
- L'axe 2 contribue largement à l'explication du taux de chômage puisque son cosinus carré est égal à 0,74.
- La variable 'indice de Gini' n'est représentée qu'à hauteur de 42% sur les 2 premières dimensions (0,29% + 0,13%), c'est en partie lié au fait que cette variable soit surtout représentative de la dimension n°3, représentant à elle seule plus de la moitié de cet axe (51% exactement).

Ainsi, réduire le plan de 9 dimensions à un plan de 2 dimensions ne permet pas de conserver une représentation parfaite de l'information initiale.

Le tableau n°10 ci-dessous est un récapitulatif de la qualité de représentation des variables explicatives. Les deux premières colonnes indiquent le degré de qualité de représentation des variables par rapport aux deux axes, et la dernière indique la qualité générale de représentation dans le cercle de corrélation à deux dimensions (1 et 2).

➤ Pour la dimension 1 on base notre étude de la 'qualité de la représentation de la variable' à la fois sur le fait que son cosinus carré soit proche ou non de 1, mais également en fonction de la moyenne. Nous allons comparer les valeurs des cosinus carrés avec la moyenne des cosinus carrés (MCC dim n°1 = 0,617), nous considérons que toute variable dont la valeur est supérieure à la moyenne est alors bien représentée par l'axe.

➤ Pour la dimension 2, étant donnée qu'elle ne représente que 13,07% de l'inertie initiale et que la moyenne de ses cosinus carrés est très faible (MCC Dim 2 = 0,13), on ne peut pas se baser sur la moyenne pour vérifier la qualité de la représentation, nous regarderons uniquement si la valeur est élevée ou non.

TABLEAU N°10 : Récapitulatif de la qualité de représentation des variables

	Dim 1 (61,5%) <i>Moyenne = 0,618</i> <i>Médiane = 0,76</i> <i>Q1 = 0,45</i> <i>Q3 = 0,83</i>	Dim 2 (13,07%) <i>Moyenne = 0,13</i> <i>Médiane = 0,01</i> <i>Q1 = 0</i> <i>Q3 = 0,13</i>	Total (74,57%)* <i>Moyenne = 0,74</i> <i>Médiane = 0,77</i> <i>Q1 = 0,45</i> <i>Q3 = 0,96</i>
--	---	--	--

PIB par habitant	bien représentée ($\cos^2 > \text{médiane}$)	pas représentée ($\cos^2 \approx 0$)	bien représentée (total $\cos^2 = 0,79$)
Salaire moyen	très bien représentée ($\cos^2 > Q3$)	pas représentée ($\cos^2 = 0$)	très bien représentée (total $\cos^2 = 0,93$)
Indice de Gini	très mal représentée ($\cos^2 < Q1$)	bien représentée ($\cos^2 = \text{moyenne}$)	très mal représentée (total $\cos^2 = 0,42$)
Indice de corruption	très bien représentée ($\cos^2 = Q3$)	pas représentée ($\cos^2 = 0$)	bien représentée (total $\cos^2 = 0,83$)
IDH	très bien représentée ($\cos^2 = Q3$)	pas représentée ($\cos^2 \approx 0$)	bien représentée (total $\cos^2 = 0,84$)
Taux de chômage	pas représentée ($\cos^2 < Q1$)	très bien représentée ($\cos^2 > Q3$)	bien représentée (total $\cos^2 = 0,78$)
Dépenses de Santé	mal représentée ($\cos^2 = Q1$)	bien représentée ($\cos^2 > Q3$)	mal représentée (total $\cos^2 = 0,72$)
Espérance de vie	bien représentée ($\cos^2 = \text{médiane}$)	pas représentée ($\cos^2 = 0$)	bien représentée (total $\cos^2 = 0,76$)
Dépenses R&D	bien représentée ($\cos^2 > \text{moyenne}$)	pas représentée ($\cos^2 \approx 0$)	mal représentée (total $\cos^2 = 0,66$)

* : Total = dim1 + dim2

Nous pouvons aussi construire un tableau reprenant les variables les plus contributives à la définition des axes, en précisant s'il s'agit d'une corrélation positive ou négative. À partir du tableau n°8 et des annexes n°4 et 5 nous créons le tableau suivant :

TABLEAU N°11 : Récapitulatif des corrélations des variables aux axes 1 et 2

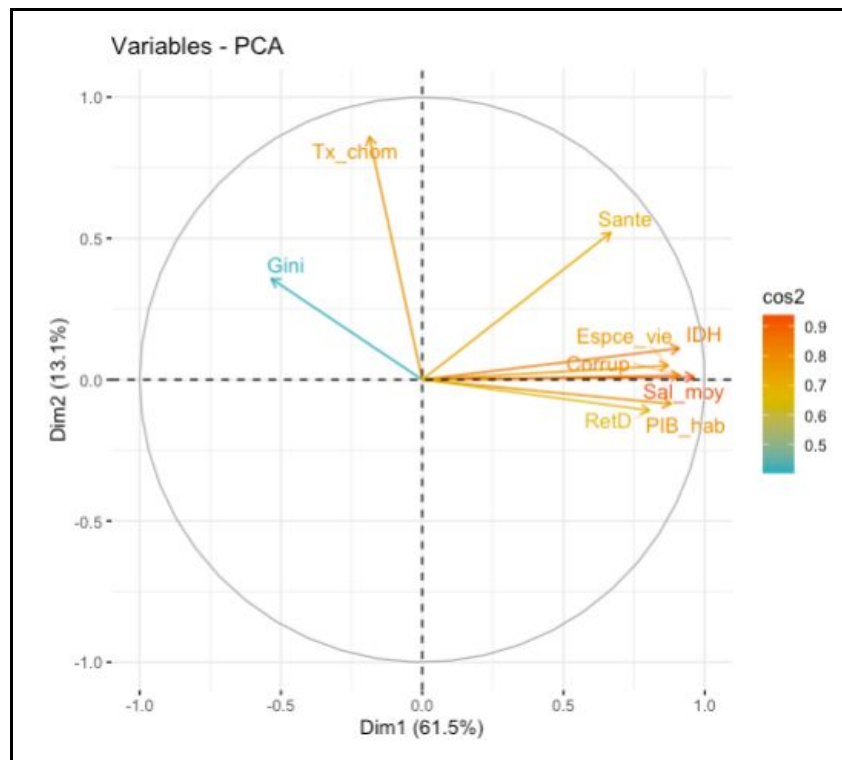
	Positif (+)	Négatif (-)
Axe n°1	Salaire moyen, niveau de corruption, IDH, PIB par habitant et espérance de vie	Indice de Gini et taux de chômage
Axe n°2	Taux de chômage, dépenses en santé et indice de Gini	PIB par habitant et R&D

C) Cercle des corrélations

Le cercle de corrélations nous donne la projection des variables initiales sur le plan des composantes principales ; les variables bien représentées dans ce plan sont celles qui sont

proches du bord du cercle, inversement celles qui sont mal représentées seront proche de l'origine.

GRAPHIQUE N°11 : Cercle de corrélation en dimension 1 et 2



➤ D'après le graphique n°11 on constate que les variables 'PIB/hab', 'salaire moyen', 'indice de corruption', 'espérance de vie', 'IDH', et 'dépenses en R&D' sont fortement corrélées entre elles, car on observe que leurs flèches sont très proches le long de l'axe 1. Elles sont toutes plus ou moins indépendantes de la variable taux de chômage car leurs flèches forment un angle droit avec celle du taux de chômage. Cette analyse du cercle de corrélation confirme les hypothèses que nous avons formulées lorsque nous avons interprété les résultats de la matrice de corrélation - ils sont donc bien bien conformes à ce que l'on a observé sur le cercle de corrélation.

➤ La variable 'dépenses de santé' est corrélée positivement avec les deux composantes principales, ses degrés de corrélation sont relativement proches comme nous pouvons le remarquer dans le tableau n°8.

➤ La variable 'taux de chômage' est fortement et positivement corrélée à l'axe 2, mais peu ou pas à l'axe 1 puisqu'on remarque qu'elle est positionnée en angle droit par rapport à celui-ci.

➤ La variable 'Indice de Gini' semble indépendante des dépenses de santé et anti-corrélée avec le 'PIB/hab' et 'dépenses en R&D'. On remarque que sa flèche est éloignée du cercle de corrélation, cela est dû au fait que comme nous l'avons vu précédemment, la somme de ses deux cosinus carrés est faible (0,42), ce qui traduit une mauvaise représentation de la variable sur le cercle de corrélation.

D) Définition des variables latentes

Chaque axe résulte de la combinaison linéaire de certaines des 9 variables. De la détermination de chaque axe va ressortir une unique variable latente qui permettra d'interpréter la position des individus dans le plan de dimensions 2. La variable latente est une variable qui numériquement n'existe pas, elle sera déterminée en fonction de plusieurs variables que nous avons identifiées dans la partie précédente.

1- Interprétation de la composante n°1

TABLEAU N°12 : Description de la dimension n°1

<code>> res.desc\$Dim.1</code>		
<code>\$quanti</code>		
	correlation	p.value
Sal_moy	0.9624175	5.885465e-34
Corrup	0.9124330	8.756666e-24
IDH	0.9104740	1.601017e-23
PIB_hab	0.8809382	3.582949e-20
Espce_vie	0.8711091	2.991380e-19
RetD	0.8046783	1.608184e-14
Sante	0.6677993	7.581288e-09
Gini	-0.5343142	1.308196e-05

En observant le cercle de corrélation et le tableau n°12, on voit que les variables qui définissent le plus l'axe n°1 (c'est à dire 'PIB/hab', 'salaire moyen', 'indice de corruption', 'espérance de vie', 'IDH', et 'dépenses en R&D') sont des variables qui caractérisent le niveau de développement d'un pays. En effet ces variables nous donnent des indications sur le niveau de revenu de la population ('PIB/hab', 'salaire moyen') et sur son état de santé ('espérance de vie', 'IDH') qui sont les critères fondamentaux de développement. On considère que les pays ayant un indice de corruption élevé (ce qui traduit une corruption faible) ont davantage tendance à être des pays développés comme nous l'avons expliqué dans notre première

partie. De même, les économies qui investissent beaucoup en R&D ont atteint un niveau de développement et de maturité supérieurs à ceux des autres pays.

On peut ainsi supposer que l'axe 1 oppose les pays sous-développés aux pays les plus développés à droite. Afin de faciliter l'analyse nous appellerons les pays à gauche de l'ordonnée les pays "peu développés" ou en "sous-développement", et ceux situés à droite les pays "développés", même si ces pays de répondent pas forcément dans la réalité aux caractéristiques des "étiquettes" qu'on leur a imposées.

2- Interprétation de la composante n°2

TABEAU N°13 : Description de la dimension 2

```
> res.desc$Dim.2
$quanti
      correlation      p.value
Tx_chom    0.8610605 2.204163e-18
Sante      0.5206163 2.369345e-05
Gini       0.3575804 5.428596e-03
```

Il paraît plus difficile d'interpréter l'axe n°2 car les corrélations des variables à celui-ci sont relativement faibles par rapport à celles de l'axe n°1. De plus, elles décroissent beaucoup dès la deuxième variable 'santé' qui y contribue à hauteur de 52%. Aussi, bien que l'indice de Gini soit assez mal représenté dans cette dimension, nous décidons tout de même de l'intégrer à cette analyse puisqu'elle apporte des informations supplémentaires qui nous seront utiles pour définir l'axe qui paraît plus compliqué à cerner.

Ainsi, nous considérerons principalement les variables 'taux de chômage', 'dépenses de santé' et 'indice de Gini' pour définir la composante n°2. Comme nous l'avons vu précédemment dans le tableau n°7 des contributions aux axes, la variable 'taux de chômage' concentre à elle seule 63,12% de la formation de l'axe 2, et les dépenses de santé 23,07%. Cela se retranscrit sur le degré de corrélation qui est plus élevé pour la variable taux de chômage (86%) que dépenses de santé (52%).

Jusqu'à présent nous n'avons pu démontrer une relation claire entre les dépenses de PIB pour la santé et le taux de pauvreté. Nous avons vu précédemment que le graphique de corrélation n'est pas concluant, que la matrice de corrélation nous indique un coefficient de corrélation légèrement négatif (-0,13), mais que le cosinus carré de la variable est trop faible pour que la variable soit considérée comme "bien représenté" sur le cercle de corrélation.

Avec les informations dont nous disposons nous ne pouvons toujours pas en conclure une relation nette et précise entre cette variable et la variable expliquée.

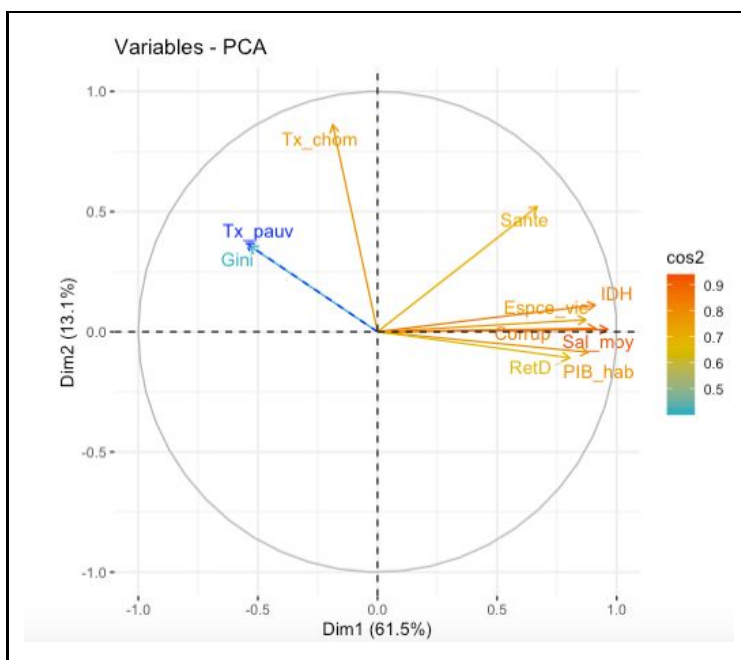
Cependant au vu du fait qu'elle contribue comme la variable taux de chômage et indice de Gini à l'axe 2, nous posons l'hypothèse qu'elle est corrélée plutôt positivement au taux de pauvreté ³² car cela a plus de sens que la relation inverse.

La variable 'taux de chômage' fait référence au niveau d'activité du pays, les dépenses en santé font référence à la part du PIB aux dépenses de santé des individus, et l'indice de Gini fait référence aux inégalités de richesse au sein d'une population. Il semble donc difficile de déterminer une variable illustrative pour l'axe 2 puisque le lien entre ces 3 variables n'est pas si évident. Cependant nous remarquons que ces 3 variables concernent les conditions de vie des individus, une personne ayant un emploi et vivant dans un pays où la population est en bonne santé et où la richesse est bien répartie entre les individus, aura de bien meilleurs conditions de vie qu'un individu au chômage dans un pays où la santé y est mauvaise et que la richesse est principalement détenue par une minorité.

On peut ainsi supposer que l'axe 2 oppose les pays aux conditions de vie convenables en bas, aux pays ayant de mauvaises conditions de vie en haut.

3- Ajout de la variable illustrative à l'ACP

GRAPHIQUE N°12 : Cercle de corrélation comprenant la variable illustrative



³² Voir justification pages 16 à 18

→ Analyse de Y par rapport aux variables

➤ Nous voyons d'après le graphique n°12 que la variable illustrative est confondue avec l'indice de Gini. Cette observation fait sens car la matrice de corrélation disponible en figure n°3, nous avait déjà révélée que la variable indice de Gini avait la corrélation la plus forte avec le taux de pauvreté (coefficient de corrélation = 0,6) : cette corrélation étant positive, il est normale que les deux flèches soient très rapprochées. De plus, nous avons remarqué dans la section précédente que la variable 'indice de Gini' était particulière puisqu'elle est corrélée négativement à l'axe 1 mais positivement au deuxième.

➤ La variable illustrative semble indépendante des dépenses de santé (elle est à son angle droit) et anti-corrélée avec le PIB par habitant et les dépenses en R&D, et dans une plus large mesure avec le regroupement entier des variables qui sont proches de l'axe 1.

→ Analyse de Y par rapport aux variables latentes

➤ La flèche représentant la variable illustrative est située en haut à gauche dans le cercle de corrélation, cela signifie que moins le pays est développé, plus il a de chance d'avoir un taux de pauvreté élevé, et moins bonnes sont les conditions de vie, plus grand sera le taux de pauvreté.

➤ La variable illustrative est corrélée aux deux axes puisque les cosinus carrés de l'angle formé par la variable illustrative aux axes 1 et 2 sont à peu près égaux à 0,5, il est néanmoins légèrement plus élevé pour l'angle formé avec l'axe n°1.

E) Projection des individus sur le plan factoriel, interprétation

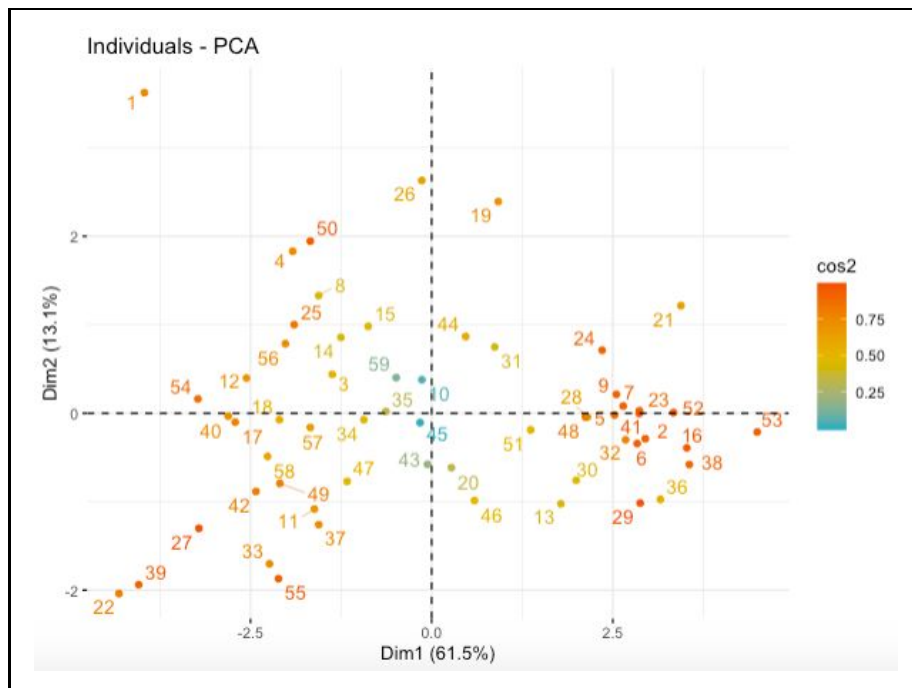
Nous allons projeter les individus sur le plan à 2 dimensions, nous pourrions ainsi voir lesquels contribuent le plus à la formation des différents axes, ce sont ceux qui sont projetés aux extrémités et proches des axes, ces individus seront les mieux représentés par les axes. À l'inverse, les individus projetés au centre seront mal représentés par les axes, ils correspondront à des individus au profil moyen, c'est-à-dire à des individus sans particularité.

Après avoir fait la projection des individus sur le plan, nous allons analyser les contributions et cosinus carrés des pays par rapport aux axes 1 et 2. Nous allons uniquement nous

intéresser aux 10 premiers pays de la base, car cela serait très fastidieux d'analyser tous les individus et cela n'apporterait pas de valeur ajoutée à l'analyse.

1- Projection des individus sur les dimensions 1 et 2

GRAPHIQUE N°13 : Projection des individus sur les axes 1 et 2



→ Analyse des pays proches du barry centre

En observant le graphique n°13 on remarque qu'il existe peu de pays situés au barry centre. Seuls le Portugal (45) et le Chili (10) sont placés à l'origine : selon l'analyse que nous avons faite des variables latentes, nous pouvons penser que ces 2 pays ne sont ni considérés développés, ni sous développés, ni avec de bonnes conditions de vie, ni avec de mauvaises. Ils sont pour ainsi dire comme la "moyenne" des composantes.

→ Analyse des pays les plus développés (partie droite du graphique)

On remarque que la majorité des pays sont regroupés le long de l'axe 1. Les pays les plus développés ont donc des conditions de vie situées au niveau de la moyenne (ni bonnes ni vraiment mauvaises). Du fait que les pays développés ne soient pas situés aux extrémités de l'axe n°2 (à l'exception de l'Espagne (19)), leur projection sur l'axe 2 est assez proche du barry centre avec une variance et un écart type plutôt faibles.

→ Analyse des pays les moins développés (partie gauche du graphique)

À l'inverse, bien que les pays les moins développés semblent avoir des conditions de vie équivalentes à celles des pays développés, ils sont davantage éparpillés dans l'espace, ce qui traduit une variance beaucoup plus élevée. De nombreux pays sont situés aux extrémités de l'axe n°2. Il est aussi curieux de noter que parmi les trois pays les moins développés, 2 d'entre eux ont des conditions de vie très bonnes (car situés en bas de l'axe 2) et l'autre a des conditions de vie mauvaises, comme le montre le tableau ci-dessous :

TABLEAU N°14 : Les 3 pays les moins développés, comparaison des conditions de vie

	Taux de chômage (en % pop)	Dépenses de santé (en % du PIB)	Indice de Gini (0<X<100)
Afrique du Sud (1)	25,4%	8,2%	62%
Pakistan (39)	3,6%	2,7%	33,5%
Ethiopie (22)	2,0%	4,0%	35%

Cela nous révèle que l'Afrique est sans doute une observation atypique.

→ Analyse des pays situés aux extrémités des axes

TABLEAU N°15 : Individus présentant une originalité quant aux axes 1 et 2

Extrémités des axes	Pays
Mauvaises conditions de vie	Afrique du Sud (1), Grèce (26), Espagne (19)
Bonnes conditions de vie	Ethiopie (22), Pakistan (39), Thaïlande (55)
Pays les plus développés	Suisse (53), Norvège (38), Danemark (16)
Pays les moins développés	Afrique de Sud (1), Ethiopie (22), Pakistan (39)

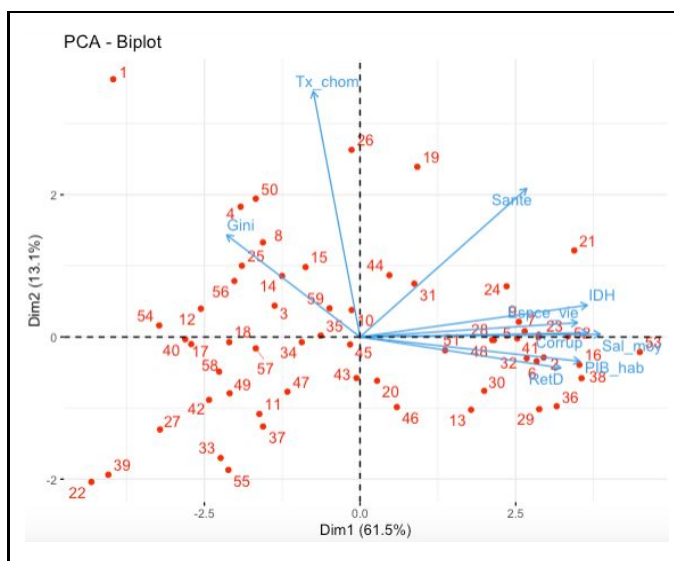
On remarque d'après le tableau n°15 que les trois pays les moins développés ont des conditions extrêmes c'est à dire bonnes pour certains et très mauvaises pour d'autres.

Le graphique n°14 ci-dessous nous donne un aperçu des relations entre les pays et les variables explicatives. On peut voir par exemple que la Grèce (26) est proche de la flèche

correspondant à la variable 'taux de chômage' ce qui signifie que ce pays a un taux de chômage particulièrement élevé en 2015, en effet il s'élève à 24,9%.

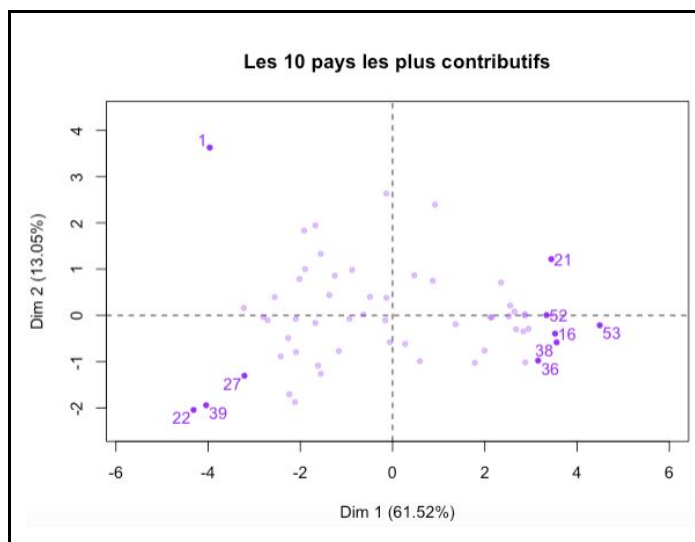
De même, le point représentant la République Tchèque (46) forme un angle à 180 degrés avec le taux de chômage nous pouvons donc supposer que son taux de chômage est particulièrement faible, il est effectivement de 5%.

GRAPHIQUE N°14 : Biplot : variables et individus en dimensions 1, 2



2- Identification des individus les plus importants dans l'explication des 2 axes

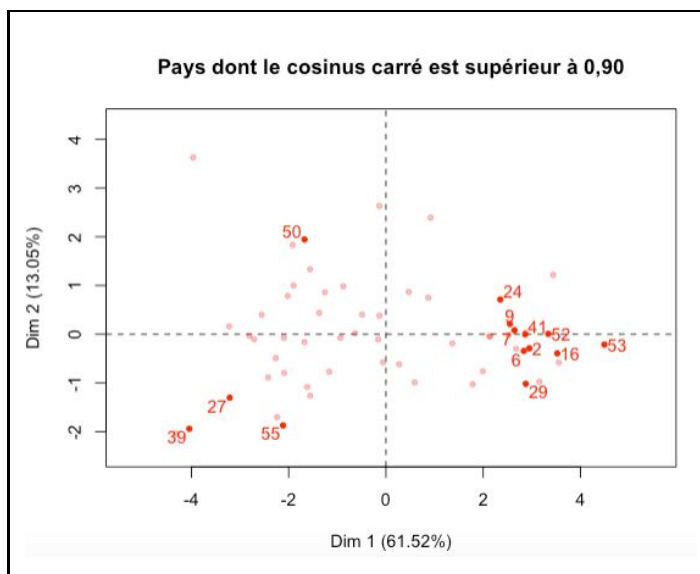
GRAPHIQUE N°15 : Top 10 des pays les plus contributifs aux dimensions 1 et 2



Le graphique n°15 et l'annexe n°6 nous donnent une indication sur les pays qui contribuent le plus à la formation des axes, c'est pourquoi ces derniers sont éloignés de l'origine. On peut néanmoins se pencher sur l'individu n°53 qui est particulièrement proche

de l'axe n°1, en revanche sa projection sur l'axe 2 est proche de zéro. On peut donc supposer que ce pays a un niveau de richesse et de développement élevé, mais qu'il n'est pas particulièrement égalitaire (indice de Gini faible) et que malgré son niveau de richesse élevé il persiste un certain taux de chômage. Il est intéressant de noter qu'il s'agit en réalité de la Suisse qui a un PIB par habitant de 82.081\$, un taux de chômage égal à 4,8% (ce qui n'est finalement pas si élevé) et un indice de Gini de 30%.

GRAPHIQUE N°16 : Pays ayant un cosinus carré supérieur à 0,9



En observant le graphique n°16 et l'annexe n°7, on voit que les pays qui sont le mieux représentés, c'est à dire ayant des cosinus carrés élevés, sont les pays les plus développés (se situant à droite du graphique), ainsi que ceux situés en bas du graphique (ceux qui ont leur représentation sur l'axe 2 la plus faible). La Serbie (50) est une exception puisqu'elle est située en haut à gauche du graphique et ne fait pas partie des pays développés alors que son cosinus carré est supérieur à 0,9, cela peut être expliqué par son taux de chômage important de 17,9% de sa population.

Nous pouvons terminer cette analyse des individus en récapitulant l'information des 4 derniers graphiques et de l'annexe n°8 en un tableau où nous indiquerons les coordonnées positives ou négatives des pays aux axes 1 puis 2.

TABLEAU N°16 : Récapitulatif des corrélations des individus aux axes 1 et 2

	Positif (+)	Négatif (-)
Axe n°1	Suisse, Norvège, Danemark et Etats-Unis	Ethiopie, Pakistan, Afrique du Sud et Tadjikistan

Axe n°2	Afrique du Sud, Grèce, Espagne et Serbie	Ethiopie, Pakistan, Thaïlande et Kazakhstan
----------------	--	---

Le tableau n°16 confirme donc notre définition des composantes principales ; si nous prenons le cas de la Norvège par exemple nous voyons que ce pays contribue fortement à la dimension 1, en effet nous remarquons qu'il a un des salaire moyens les plus élevés ; 50.994\$ par habitant et par an. Nous pouvons aussi prendre l'exemple de l'Ethiopie qui contribue négativement aux 2 axes : nous attendons donc un niveau de développement assez faible que nous vérifierons en regardant le salaire moyen et le taux de corruption de ce pays, et de bonnes conditions de vie ; nous irons cette fois regarder son taux de chômage et les dépenses en santé en pourcentage de son PIB.

Dans notre base nous voyons que son salaire moyen est extrêmement faible puisqu'il est de 478,7\$ par tête et par an. En ce qui concerne la corruption l'indice est de 33%, sachant que plus il est proche de 0 plus grande est la corruption, ce constat confirme l'interprétation de la dimension 1 : l'Ethiopie est un pays sous développés.

De plus, dans la partie économique nous avons pu, grâce aux graphique de corrélation, classer les continents en sous groupes selon leur niveau de développement. Nous avons ainsi classé l'Afrique en "défavorisé" cela rejoint donc ce que nous venons d'affirmer concernant l'Ethiopie qui fait partie de ce continent.

Partie 4 : Corrélation entre la variable à expliquer et les variables latentes

A) Régression linéaire multiple

Nous allons essayer de comprendre la variation du taux de pauvreté dans une logique de régression. Grâce à l'ACP réalisée précédemment nous avons identifié des variables latentes, nous pouvons désormais introduire le taux de pauvreté qui était jusqu'à présent une variable illustrative, c'est-à-dire qu'elle ne participait pas à la définition des facteurs.

Nous allons commencer par faire une régression linéaire multiple, notée RLM par la suite, en utilisant la méthode des Moindres Carrés Ordinaires. Grâce à cette dernière, nous allons voir si le taux de pauvreté est corrélé positivement ou négativement avec chaque axe. Nous procéderons donc à l'analyse en 3 étapes : d'abord nous regarderons la significativité des variables latentes pour n'interpréter que celles qui ont un impact sur le taux de pauvreté, sinon elles ne sont pas intéressantes d'un point de vue économétrique, ensuite nous regarderons le signe des coefficients estimés par la RLM, et enfin nous analyserons les coefficients eux-mêmes.

1- Estimation des coefficients

TABLEAU N°17 : Régression linéaire multiple sur les 2 axes

```
> summary(lm)

Call:
lm(formula = basepauvretéTx_pauv ~ res.pca$ind$coord[, 1] +
    res.pca$ind$coord[, 2])

Residuals:
    Min       1Q   Median       3Q      Max
-14.2082  -3.8300  -0.2559   4.4225  11.5487

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    15.1153     0.7716  19.591 < 2e-16 ***
res.pca$ind$coord[, 1] -1.8065     0.3279  -5.509 9.41e-07 ***
res.pca$ind$coord[, 2]  2.6350     0.7119   3.701 0.000491 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.926 on 56 degrees of freedom
Multiple R-squared:  0.4403,    Adjusted R-squared:  0.4203
F-statistic: 22.03 on 2 and 56 DF,  p-value: 8.763e-08
```

D'après le tableau n°17, on peut voir que les 2 axes sont significatifs au seuil de risque de 0,1%, nous pourrions donc interpréter le signe et les coefficients des 2 composantes.

➤ Le coefficient associé à l'axe 1 est négatif, étant donné que la variable latente correspondante à l'axe 1 est le niveau de développement, cela fait sens qu'elle soit corrélée négativement au taux de pauvreté du pays. En effet, cela signifie que plus un pays est développé moins il y aura de pauvreté sur son territoire, toutes choses égales par ailleurs. Le coefficient associé est égale à -1,81, cela veut dire que quand le niveau de développement augmente d'un point alors le taux de pauvreté diminue de 1,81%, l'impact est donc plus que proportionnel ($1,81 > 1\%$).

➤ Le coefficient du deuxième axe est quant à lui positif, sachant que la variable latente correspondante est les conditions de vie dans un pays, cela semble aussi logique. Cette variable latente est principalement définie par le taux de chômage : lorsque les conditions de vie s'améliorent alors le taux de pauvreté baisse dans le pays concerné puisque les conditions de vie deviennent mauvaises en montant le long de l'axe n°2. Le coefficient de corrélation est de 2,64, cela signifie que lorsque les conditions de vie s'empirent d'1 point, le taux de pauvreté augmentera de 2,63%. La pauvreté est donc très sensible à cette variable, principalement définie par le taux de chômage et les dépenses en santé. Ici, l'impact d'une variation de l'axe est encore plus importante que pour l'axe n°1 ($2,63 > 1,81 > 1\%$).

➤ Concernant le R^2 nous constatons qu'il est très faible : la RLM appliquée aux 2 premières composantes ne permet d'expliquer que 45% du taux de pauvreté.

2- Analyse des résidus

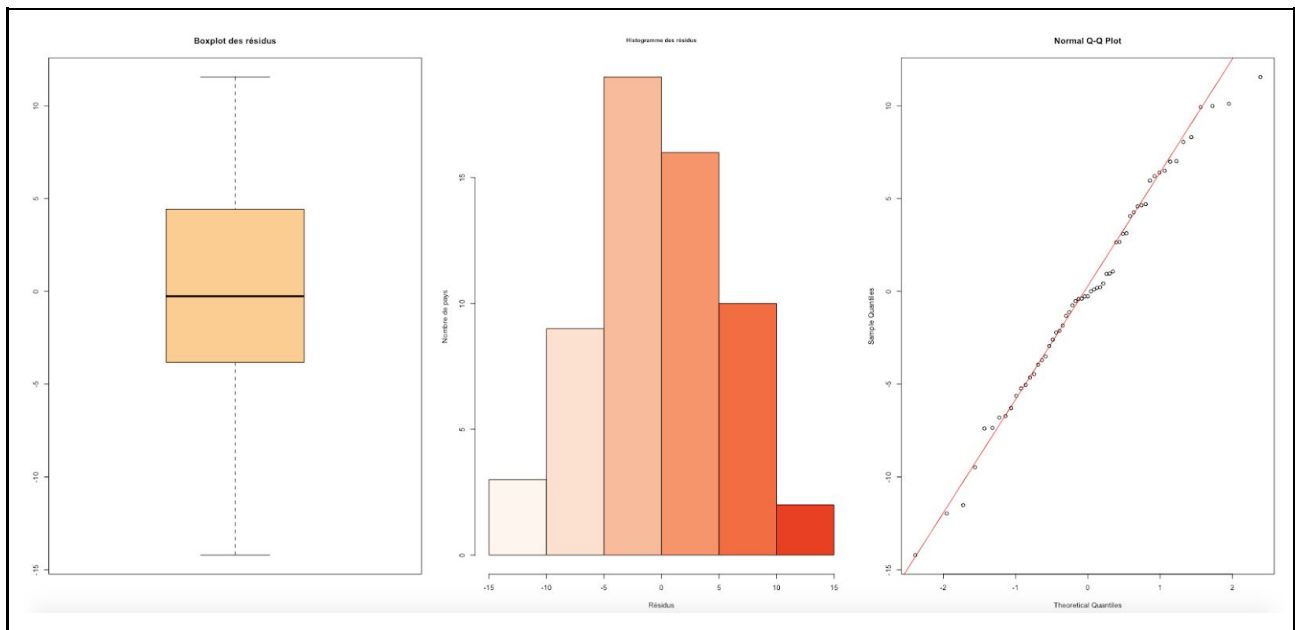
Nous allons désormais vérifier si les résidus du modèle sont distribués selon une loi normale, c'est une hypothèse forte pour que le modèle soit considéré de bonne qualité. Nous allons pouvoir constater si cette hypothèse est vérifiée à l'aide de coefficients, d'analyses graphiques, ainsi que de tests.

Commençons par vérifier les coefficients de Skewness et de Kurtosis pour la distribution des résidus. Comme vu précédemment pour la distribution des variables actives, les coefficients de symétrie (skewness) et d'aplatissement (kurtosis) prennent la valeur 0 lorsque les valeurs suivent une loi normale. D'après le calcul sous R, le coefficient de Skewness prend la valeur -0,144 → le fait qu'il soit négatif indique que la distribution des résidus est étalée à gauche donc asymétrique à droite, mais est néanmoins proche de la symétrie parfaite d'une loi normale. Le coefficient de Kurtosis quant à lui est de 2,636, il est également différent

de 0 \rightarrow le fait qu'il soit positif indique que la distribution des résidus est moins aplatie qu'une distribution suivant une loi normale.

Passons désormais à l'analyse graphique des résidus, nous allons utiliser pour cela une représentation de la distribution sous forme d'histogramme, de boîte à moustache et de diagramme quantile-quantile noté Q-Q par la suite.

FIGURE N°4 : Boxplot, histogramme et diagramme Q-Q des résidus



D'après la figure n°4, nous constatons en regardant l'histogramme qu'effectivement la distribution n'est pas symétrique mais légèrement étalée à gauche. Pour analyser la boîte à moustaches, il faut tout d'abord savoir à quoi correspondent les limites de la boîte ainsi que les moustaches. La ligne noire au milieu de la boîte correspond à la médiane, les bords de la boîte correspondent aux quantiles 1 et 3, et les moustaches correspondent dans notre cas aux valeurs extrêmes des résidus c'est à dire au minimum et au maximum (nous aurions pu changer cela en ajoutant le paramètre graphique "range" différent de 0, mais nous décidons de le laisser ainsi pour englober toutes les valeurs des résidus). Nous constatons donc que la médiane ne se situe pas parfaitement au milieu de la boîte, nous pourrions l'estimer à -0,1, donc la distribution n'est pas symétrique. Enfin, nous remarquons d'après le diagramme Q-Q que certains points ne se situent pas sur la droite de Henry, donc les résidus ne semblent pas suivre une loi normale.

Pour conclure nous pouvons dire que la distribution des résidus ne s'apparente pas à une loi normale, cependant elle en est très proche car la médiane est presque égale à 0 sur le boxplot, visuellement nous ne sommes pas très loin de la distribution normale si l'on regarde l'histogramme, enfin, les points sont presque alignés sur la droite (QQ plot) ce qui traduit une distribution proche de la loi normale.

On peut à présent confirmer ou invalider ce doute en effectuant un test statistique qui nous indiquera si les résidus suivent ou non une loi normale. Etant dans le doute quant à la distribution suivie par les résidus, il nous faudra chercher si notre modèle valide ou non les hypothèses requises pour appliquer un test non paramétrique : c'est la sortie du "Global Validation of Linear Models Assumptions" (GVLMA) sous R qui nous apportera cette réponse. Si les hypothèses ne sont pas satisfaites nous déciderons d'appliquer un test non paramétrique qui est le test général qui détermine la normalité des résidus. Ce dernier présente l'avantage qu'il peut être utilisé dans un plus grand nombre de situations puisqu'il ne nécessite aucune hypothèse de validité. En revanche, étant fait pour des cas non définis et plus larges, il sera moins précis qu'un test paramétrique qui lui requiert un certain nombre d'hypothèses pour être appliqué. L'une d'entre elles concerne la distribution des variables et des résidus ; elle se doit d'être normale. Nous ne sommes pas sûrs ni de la distribution normale des individus, ni de celle des variables c'est pourquoi nous allons d'abord appliquer la commande GVLMA pour savoir quel test utiliser.

TABLEAU N°18 : Évaluation générale de la qualité du modèle

```
> gvlma(lm)

Call:
lm(formula = basepauvreté$Tx_pauv ~ res.pca$ind$coord[, 1] +
    res.pca$ind$coord[, 2])

Coefficients:
      (Intercept)  res.pca$ind$coord[, 1]  res.pca$ind$coord[, 2]
           15.115                -1.807                 2.635

ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance = 0.05

Call:
gvlma(x = lm)

      Value p-value      Decision
Global Stat    1.2147 0.8757 Assumptions acceptable.
Skewness       0.2044 0.6512 Assumptions acceptable.
Kurtosis       0.3254 0.5684 Assumptions acceptable.
Link Function  0.2079 0.6485 Assumptions acceptable.
Heteroscedasticity 0.4770 0.4898 Assumptions acceptable.
```

Nous constatons à partir du tableau n°18 que toutes les hypothèses sont validées : la situation s'éclaircit, nous appliquerons un test non paramétrique dans un premier temps, puis un test paramétrique afin de comparer leurs résultats et voir lequel convient le mieux.

Nous appliquerons d'abord le test paramétrique de Kolmogorov Smirnov. Celui-ci est valable pour les échantillons supérieurs à 50 observations comme c'est le cas dans notre étude. Il permet d'évaluer la normalité des résidus du modèle selon les hypothèses suivantes:

H_0 : les résidus suivent une loi normale

H_1 : les résidus ne suivent pas une loi normale

TABLEAU N°19 : Résultats du test de Kolmogorov-Smirnov

```
One-sample Kolmogorov-Smirnov test

data: basepauvreté$residu
D = 0.07119, p-value = 0.9051
alternative hypothesis: two-sided
```

Nous constatons d'après le tableau n°19 que la p-value associée au test est de 0,91, elle est largement supérieure à 0,05 par conséquent nous acceptons l'hypothèse H_0 selon laquelle les résidus suivent une loi normale au seuil de risque de 5%.

Nous pouvons à présent comparer ce résultat avec le test non paramétrique donc plus précis de Levene. Les hypothèses sont les mêmes, après avoir séparé notre échantillon en 2, à partir du logiciel R nous obtenons :

TABLEAU N°20 : Résultats du test de Levene

```
> leveneTest(y,groupe)
Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group 1  0.0526 0.8195
    57
```

D'après le tableau n°20 nous voyons que la p-value associée au test est de 0,82, encore une fois nous acceptons l'hypothèse H_0 ; les résidus suivent donc une loi normale au seuil de risque de 5%.

B) Relation entre la variable à expliquer observée et prédite par RLM

Dans cette nouvelle partie nous allons comparer les valeurs que nous avons récoltées pour le taux de pauvreté en 2015 avec celle prédites par le modèle précédent. Cela traduira donc la pertinence de notre modèle et sa proximité avec la réalité si le taux de pauvreté prédit est proche du taux de pauvreté observé.

TABLEAU N°21 : Taux de pauvreté prédit et taux de pauvreté observé

```
> round(basepauvreté$predict,2)
[1] 31.83 9.03 18.75 23.40 10.50 9.09 10.55 21.44 11.07 16.36 15.18 20.78 9.19 19.64 19.28 7.72 19.75 18.72 19.76 13.00 12.10
[22] 17.53 10.01 12.74 21.18 22.29 17.49 11.12 7.24 9.51 15.52 9.49 14.67 16.61 16.31 6.84 7.16 17.31 20.10 9.93 17.16
[43] 13.70 16.55 15.12 11.44 15.19 11.17 16.81 23.26 12.15 9.10 6.44 21.37 14.00 20.83 17.72 17.92 17.06
> basepauvreté$Tx_pauv
[1] 26.6 10.1 30.3 29.8 10.5 8.7 9.8 25.7 14.2 16.1 5.7 27.8 17.5 20.6 19.5 5.5 27.8 23.3 15.3 16.1 16.8 23.5 6.3 8.1 21.6
[26] 14.9 11.2 9.8 5.4 19.5 14.4 15.7 2.7 16.2 16.5 10.9 0.4 8.1 24.3 26.6 7.8 21.8 11.1 12.6 8.4 6.4 25.3 10.9 13.3 25.9
[51] 9.2 9.2 9.1 31.3 7.2 15.2 17.2 6.4 9.7
> basepauvreté$residu
[1] -5.23 1.07 11.55 6.40 0.00 -0.39 -0.75 4.26 3.13 -0.26 -9.48 7.02 8.31 0.96 0.22 -2.22 8.05 4.58
[19] -4.46 3.10 4.70 5.97 -3.71 -4.64 0.42 -7.39 -6.29 -1.32 -1.84 9.99 -1.12 6.21 -11.97 -0.41 0.19 4.06
[37] -14.21 0.94 6.99 6.50 -2.13 4.64 -2.60 -3.95 -6.72 -5.04 10.11 -0.27 -3.51 2.64 -2.95 0.10 2.66 9.93
[55] -6.80 -5.63 -0.52 -11.52 -7.36
```

La sortie de la commande “basepauvreté\$residu” nous donne la valeur des résidus de Y associé à chaque observation. Ainsi si elle est grande cela signifie que le taux de pauvreté prédit est très différent de celui observé, donc que le modèle ne retranscrit pas bien les variations du taux de pauvreté entre les pays. Au contraire, si la valeur est faible alors la prédiction du modèle sera bonne voire très bonne.

À partir du tableau n°21, on constate des prévisions assez différentes puisqu’elles vont de 0 (excellente prévision du modèle) à -14,21 : dans ce cas les valeurs prédites de Y sont supérieures aux valeurs observées puisque les résidus sont calculés de la manière suivant $u=Y-\hat{Y}$ où Y est le taux de pauvreté observé (“basepauvreté\$Tx_pauv”) et \hat{Y} est le taux de pauvreté prédit par notre modèle (“round(basepauvreté\$predict,2)”).

On voit par exemple que la prévision est excellente pour le 5ème individu pour lequel le taux de pauvreté observé est égal au taux prédit, l’estimation est donc parfaite : nous nous attendons à le voir sur la courbe du graphique ayant en abscisse le taux prédit et en ordonnées le taux observé. On note cependant des oppositions puisque pour l’individu n°37 la prévision est très mauvaise : $Y-\hat{Y}=-14,21$: nous nous attendons donc à le voir en dessous et éloigné de la droite du modèle sur le graphique.

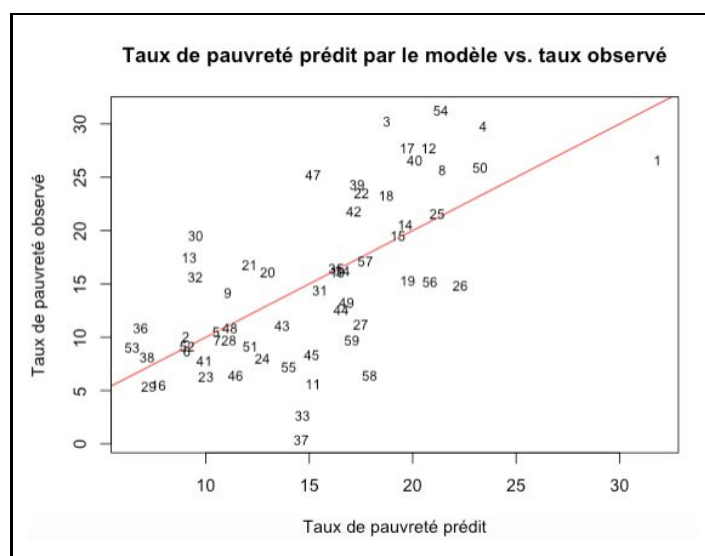
TABLEAU N°22 : Tableau récapitulatif des pays les mieux et les plus mal représentés par le modèle

Pays dont le taux de pauvreté est correctement estimé par le modèle (résidus < 1)	Pays dont le taux de pauvreté est très mal estimé par le modèle (résidus > 8)
Australie (5), Autriche (6), Belgique (7), Chili (10), Costa Rica (14), Croatie (15), Géorgie (25), Lettonie (34), Lituanie (35), Norvège (38), Royaume-Uni (48), Suède (52), Turquie (57)	Argentine (3), Chine (11), Corée (13), Egypte (17), Israël (30), Kazakhstan (33), Malaisie (37), Roumanie (47), Tadjikistan (54), Ukraine (58)

D'après le tableau n°22, on peut effectuer plusieurs remarques :

- Les pays de la colonne de gauche seront normalement situés très proches de la droite de régression sur le graphique représentant le modèle. A l'inverse les pays de la colonne de droite seront plus éloignés de cette droite.
- On remarque que les pays les mieux représentés sont généralement des pays d'Europe, à l'exception du Chili et du Costa Rica, alors que dans les pays les moins bien représentés seule l'Ukraine fait partie de l'Europe.
- On note également que parmi les pays les mieux représentés, il y a une parité entre les pays considérés comme développés (à droite du graphique) et ceux les moins développés (à gauche du graphique).³³ Alors que du côté de ceux qui sont mal représentés, seul la Corée est un pays considéré comme développé. Il apparaît donc que les pays développés sont mieux représentés par le modèle que les pays moins développés.

GRAPHIQUE N°17 : Prédiction du taux de pauvreté par la régression



³³ Voir page 41 , graphique n°13 : Projection des individus sur les axes 1 et 2

La droite rouge représente la droite de régression du modèle, ainsi tout individu se situant sous celle-ci a un taux de pauvreté prédit supérieur au taux de pauvreté observé. Comme nous l'avons prédit précédemment, les individus se trouvant sur la droite sont les mêmes que ceux situés dans la colonne de gauche, d'après le graphique n°17 on peut retrouver les numéros 25, 14, 15, et 57 par exemple. À l'inverse, les plus éloignés sont les numéros 37, 33, 3, et 47 par exemple. Ainsi les conclusions émises à partir du tableau n°22 sont validées et confirmées par le graphique n°17.

C) Régression logistique binaire

Nous allons maintenant transformer la nature de notre variable à expliquer en la passant de variable quantitative continue à variable binaire. La moyenne du taux de pauvreté sur notre échantillon étant d'environ 15,12% nous faisons le choix de séparer les pays en deux d'après ce critère. Notre variable qualitative prend la valeur "1" si le taux de pauvreté du pays du pays est strictement supérieur à 15,12% et "0" sinon (ou si égal à 15,12%).

TABLEAU N°23 : Modèle logistique binaire, codage de Y par rapport à sa moyenne

```
> summary(modele)

Call:
glm(formula = basepauvretéTx_pauvbiN ~ res.pca$ind$coord[,
  1] + res.pca$ind$coord[, 2], family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0777  -0.7405  -0.3605   0.7665   2.0451

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.1317     0.3251  -0.405 0.685421
res.pca$ind$coord[, 1] -0.5684     0.1591  -3.573 0.000353 ***
res.pca$ind$coord[, 2]  0.7945     0.3471   2.289 0.022091 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 81.639  on 58  degrees of freedom
Residual deviance: 59.701  on 56  degrees of freedom
AIC: 65.701

Number of Fisher Scoring iterations: 4
```

D'après le tableau n°23, les coefficients associés aux deux axes sont significatifs au seuil de 5%, nous les appelons *log(pile)* et *log(face)*. Le coefficient associé à l'axe 1 est négatif, cela veut dire que les caractéristiques de l'axe 1 réduisent la probabilité pour un pays d'avoir un taux de pauvreté élevé. Pour pouvoir interpréter librement le coefficient estimé de chaque composante lorsque Y est codée de manière binaire, il faut procéder à la "transformation" de

la forme logarithmique $\rightarrow \log(pile) = -0,5684 \Leftrightarrow -(\exp(-0,5684)) = -0,5664$.³⁴ Ainsi, la probabilité pour un pays d'avoir un taux de pauvreté supérieur à 15,12%, sera 0,57 fois inférieure pour un pays développé.

Nous faisons de même pour le coefficient de l'axe n°2 : $\log(face) = 0,7945 \Leftrightarrow \exp(0,7945) = 2,21$ Ainsi la probabilité pour un pays d'avoir un taux de pauvreté supérieur à 15,12% , sera 2,21 fois supérieure pour un pays ayant de mauvaises conditions de vie.

Néanmoins il est important de noter que la p-value associée à l'axe 2 est tout juste inférieure à 0,05 (p-value=0,02), nous voudrions que cet axe soit significatif au seuil de 1% c'est pourquoi nous allons effectuer la même analyse mais en codant Y par rapport à sa médiane qui est de 14,4% pour voir si cela augmente la significativité de l'axe 2.

TABLEAU N°24 : Modèle logistique binaire, codage de Y par rapport à sa médiane

```
> summary(modele1)

Call:
glm(formula = basepauvretéTx_pauvterN ~ res.pca$ind$coord[,
  1] + res.pca$ind$coord[, 2], family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5678  -0.6827  -0.2927   0.6261   2.1968

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.03752    0.34931   0.107  0.914466
res.pca$ind$coord[, 1] -0.63261    0.17297  -3.657  0.000255 ***
res.pca$ind$coord[, 2]  1.19901    0.41863   2.864  0.004181 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 81.774  on 58  degrees of freedom
Residual deviance: 54.474  on 56  degrees of freedom
AIC: 60.474

Number of Fisher Scoring iterations: 5
```

D'après le tableau n°24, on remarque que le coefficient associé à l'axe 1 est plus élevé dans ce cas ; donc la probabilité pour un pays d'avoir un taux de pauvreté supérieur à 14,4%, sera de $-(\exp(-0,632))=0,53$ fois inférieure pour un pays développé.

Le coefficient associé à l'axe 2 est significatif au seuil de 1% comme nous voulions, en revanche il est beaucoup plus élevé que lors de la première régression donc sera beaucoup plus faible lors de la retranscription en exponentiel. Ainsi, la probabilité pour un pays d'avoir

³⁴ On a passé l'expression sous forme négative car la fonction exponentielle est toujours positive pour tout x

un taux de pauvreté supérieur à 14,4%, sera $\exp(1,199)=3,32$ fois supérieure pour un pays ayant de mauvaises conditions de vie.

Nous pouvons terminer cette partie en récapitulant les informations du modèle RLM, significativité des axes, distribution des résidus ...

TABLEAU N°25 : Résumé des caractéristiques du modèle de RLM nommé "lm"

	Modèle n°1 "lm"	Conclusion
R²	0,44	✗*
R² ajusté	0,42	✗
Nombre de variables significatives	2/2	✓
Boxplot des résidus	médiane très proche de 0	✓
Histogramme des résidus	légère asymétrie à droite	✗
QQ plot des résidus	points presque alignés	✓
Skewness et kurtosis	-0.144 et 2.636	✗
Qualité du modèle : GVLMA	Hypothèses acceptées	✓
Test paramétrique	p-value=0,91	✓
Test non paramétrique	p-value=0,82	✓

* : Les croix veulent seulement dire que la situation n'est pas optimale pour la condition concernée, en revanche la pertinence du modèle reste bonne dans l'ensemble.

Partie 5 : Conclusion et discussion des résultats

Pour conclure notre analyse nous allons revenir sur certains points que nous avons évoqués au cours des explications :

Il est important de rappeler que l'étude des deux premières dimensions permet de conserver 74,5 % des caractéristiques des individus, nous pouvons donc dire que les analyses réalisées à partir de l'ACP représentent assez fidèlement la réalité. Le niveau de développement est l'axe le plus important car il représente à lui seul 61,5% de l'inertie totale, c'est cet axe qui ressort comme étant le plus significatif.

De plus, notre analyse semble relativement solide, cependant elle pourrait être améliorée. Il aurait été intéressant d'observer une variable concernant les diplômés de l'enseignement supérieur, ou encore une variable représentative du système de redistribution des richesses, via les revenus de transfert venant de l'Etat comme nous l'avons vu en introduction. Cela aurait peut-être permis d'améliorer le coefficient de détermination et de trouver de nouveaux axes de réflexion.

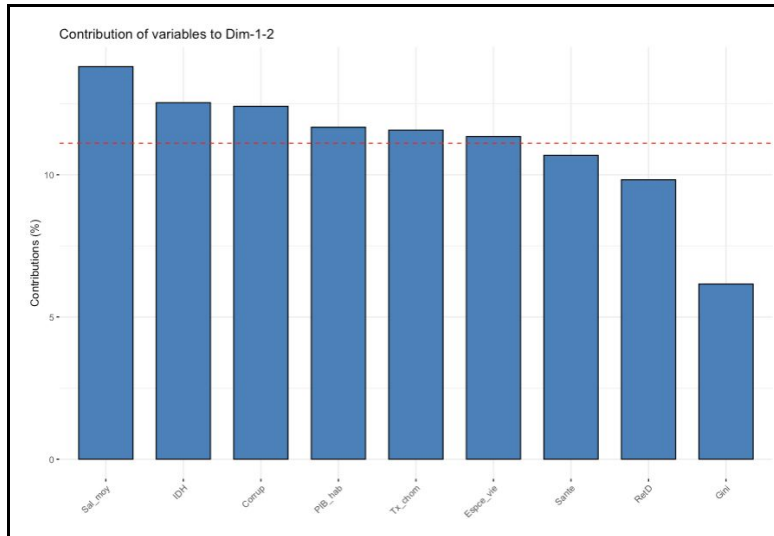
Une critique peut être émise quant à la qualité de notre modèle, en effet les coefficients de Skewness et de Kurtosis indiquent que les résidus ne semblent pas être distribués selon une loi normale. Ces résultats sont contredits par les tests de Kolmogorov et de Levene, nous nous questionnons donc sur la fiabilité des coefficients de symétrie et d'aplatissement.

Enfin avec les informations dont nous disposons grâce à notre analyse (graphique et matrice de corrélation, cosinus carré et répartition des pays dans le cercle de corrélation), nous n'avons pas été capable déterminer une relation claire entre la variable 'dépenses de santé' et la variable à expliquer. Afin de déterminer la variable latente de l'axe n°2, nous avons donc émis l'hypothèse qu'elle était corrélée plutôt positivement au taux de pauvreté. Nous tenons à dire que si nous avions eu un autre échantillon de pays ou à une autre période, la relation entre les deux variables aurait pu être totalement différente de celle de notre hypothèse, ici nous nous sommes fié uniquement à notre analyse et aux données dont nous disposions.

Partie 6 : Annexes

A) Tableaux supplémentaire utilisés dans l'analyse

Annexe n°1 : Contribution de chaque variable aux axes 1 et 2

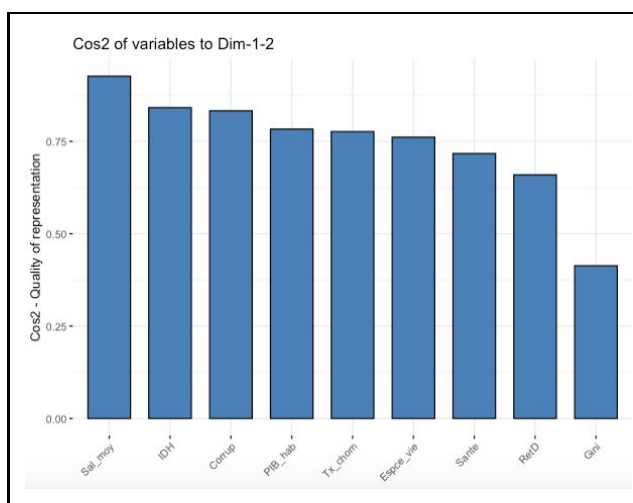


Annexe n°2 : Tableau cosinus carré de chaque variable par rapport à chaque dimension

```
> round(res.pca$var$cos2,2)
```

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
PIB_hab	0.78	0.01	0.00	0.08	0.07
Sal_moy	0.93	0.00	0.00	0.01	0.02
Gini	0.29	0.13	0.51	0.04	0.01
Corrup	0.83	0.00	0.00	0.02	0.03
IDH	0.83	0.01	0.00	0.01	0.07
Tx_chom	0.04	0.74	0.18	0.00	0.00
Sante	0.45	0.27	0.07	0.09	0.02
Espce_vie	0.76	0.00	0.00	0.01	0.18
RetD	0.65	0.01	0.02	0.20	0.00

Annexe n°3 : Histogramme de la somme des cos2 de chaque variable aux axes 1 et 2



Annexe n°4 : Résumé des variables les plus corrélées aux 2 axes

```
> dimdesc(res.pca)
$Dim.1
$Dim.1$quanti
      correlation      p.value
Sal_moy    0.9624175 5.885465e-34
Corrup     0.9124330 8.756666e-24
IDH        0.9104740 1.601017e-23
PIB_hab    0.8809382 3.582949e-20
Espce_vie  0.8711091 2.991380e-19
RetD       0.8046783 1.608184e-14
Sante      0.6677993 7.581288e-09
Gini       -0.5343142 1.308196e-05

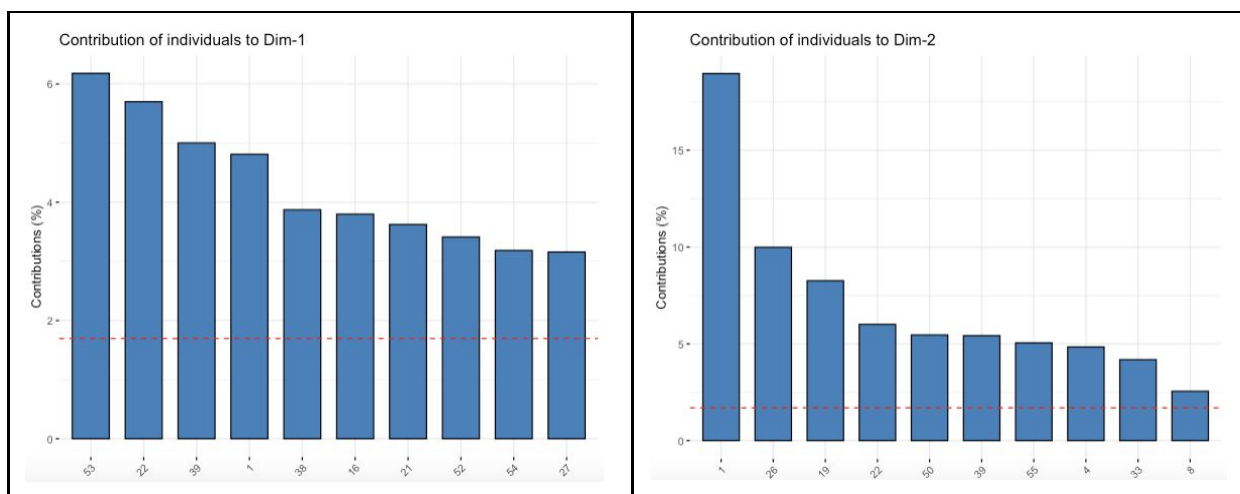
$Dim.2
$Dim.2$quanti
      correlation      p.value
Tx_chom    0.8610605 2.204163e-18
Sante      0.5206163 2.369345e-05
Gini       0.3575804 5.428596e-03

$Dim.3
$Dim.3$quanti
      correlation      p.value
Gini       0.7171591 1.662779e-10
Sante      0.2664287 4.137816e-02
Tx_chom    -0.4281769 7.168948e-04
```

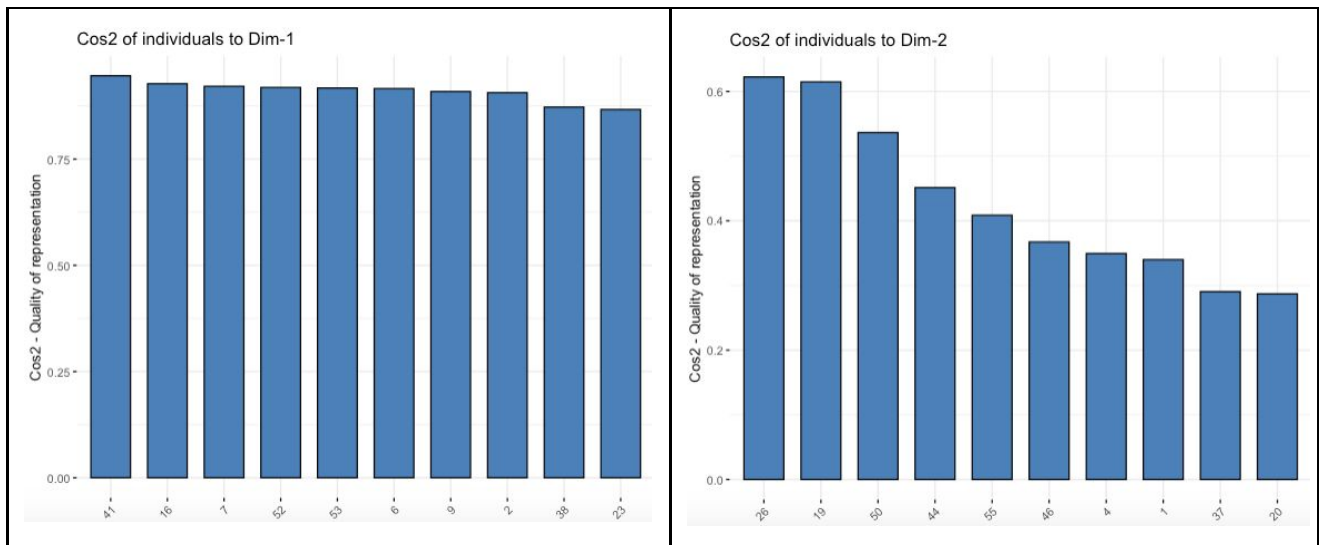
Annexe n°5 : Corrélations des variables aux axes 1 et 2

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
PIB_hab	0.88	-0.09	0.06	0.28	-0.26
Sal_moy	0.96	0.01	0.01	0.09	-0.15
Gini	-0.53	0.36	0.72	0.19	0.10
Corrup	0.91	0.01	0.05	0.15	-0.16
IDH	0.91	0.11	-0.07	0.11	0.27
Tx_chom	-0.19	0.86	-0.43	0.04	-0.04
Sante	0.67	0.52	0.27	-0.29	-0.13
Espce_vie	0.87	0.05	-0.03	0.12	0.42
RetD	0.80	-0.11	0.12	-0.45	0.05

Annexe n°6 : Top 10 des pays ayant la plus forte contribution aux axes 1 et 2



Annexe n°7 : Top 10 des pays ayant le plus fort cosinus carré par rapport aux axes 1 et 2



Annexe n°8 : Coordonnées des individus aux axes 1 et 2

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
1	-3.96	3.63	1.44	0.14	-1.79
2	2.95	-0.29	0.31	-0.72	-0.21
3	-1.37	0.44	0.76	0.04	0.53
4	-1.92	1.83	-1.12	-0.66	-0.25
5	2.52	-0.02	0.58	0.78	0.05
6	2.84	-0.34	0.05	-0.71	-0.26
7	2.64	0.08	-0.40	-0.44	-0.31
8	-1.56	1.33	2.12	-0.71	0.14
9	2.55	0.21	0.22	0.34	-0.16
10	-0.13	0.38	1.18	0.98	0.58
11	-1.62	-1.08	0.46	-0.69	0.60
12	-2.56	0.40	1.36	0.73	0.68
13	1.78	-1.02	0.64	-1.25	1.16
14	-1.25	0.86	1.13	0.72	0.78
15	-0.88	0.98	-1.34	0.03	0.33
16	3.52	-0.39	-0.17	-0.41	-0.63
17	-2.71	-0.10	-1.22	-0.29	-0.43
18	-2.10	-0.07	1.52	-0.06	0.49
19	0.92	2.39	-1.30	0.48	0.53
20	0.27	-0.62	-0.17	0.15	0.30
21	3.44	1.22	1.98	-0.88	-1.16
22	-4.32	-2.04	0.27	-0.86	-1.98
23	2.87	0.03	-0.57	-0.48	-0.31
24	2.35	0.71	-0.29	-0.57	0.04
25	-1.90	1.00	-0.51	-0.03	-0.27
26	-0.14	2.63	-1.74	0.35	0.69
27	-3.21	-1.30	0.18	0.45	-0.15
28	2.15	-0.04	-0.68	1.38	-0.27
29	2.88	-1.02	-0.37	0.35	-0.10
30	1.99	-0.76	0.67	-1.05	0.93

31	0.87	0.75	-0.50	0.18	0.73
32	2.68	-0.30	0.94	-0.93	0.59
33	-2.24	-1.70	-1.33	0.24	0.10
34	-0.94	-0.07	-0.50	0.55	-0.03
35	-0.63	0.02	-0.07	0.31	0.00
36	3.16	-0.98	-0.32	2.55	-1.09
37	-1.56	-1.26	0.65	0.29	0.62
38	3.56	-0.58	-0.04	0.78	-0.65
39	-4.04	-1.94	-0.34	-0.22	-1.26
40	-2.81	-0.03	1.50	0.11	0.13
41	2.87	0.00	-0.09	0.17	-0.31
42	-2.42	-0.88	0.88	0.62	0.60
43	-0.06	-0.58	-0.81	0.15	0.22
44	0.47	0.87	-0.39	0.10	0.45
45	-0.16	-0.11	-1.55	-0.27	0.07
46	0.59	-0.99	-0.73	-0.59	0.50
47	-1.17	-0.77	-0.17	0.99	-0.18
48	2.12	-0.04	0.70	0.54	-0.10
49	-2.09	-0.79	0.15	-0.28	0.15
50	-1.67	1.94	-0.32	-0.47	0.14
51	1.37	-0.19	-1.01	-0.71	0.48
52	3.34	0.01	0.03	-0.69	-0.28
53	4.49	-0.21	0.67	-0.10	-0.85
54	-3.23	0.16	-0.64	-0.51	-0.72
55	-2.11	-1.87	0.21	0.25	0.65
56	-2.02	0.79	-1.09	-0.31	0.14
57	-1.67	-0.16	-0.16	0.57	0.77
58	-2.26	-0.49	-1.34	-0.96	-0.34
59	-0.49	0.40	0.69	0.51	-0.10

B) Script R studio

On importe la base puis on renomme les variables

```
library(readxl)
basepauvreté=read_excel(file.choose(),col_names=T)
basepauvreté
View(basepauvreté)
```

```
install.packages("questionr",dependencies=TRUE)
install.packages("FactoMineR",dependencies=TRUE)
install.packages("factoextra")
install.packages("moments")
install.packages("gvlma")
install.packages("moments")
install.packages("car")
```

I) Nettoyage de la base !!

#On renomme les variables par des noms plus courts mais explicites

```
library(questionr)
```

```

names(basepauvreté)
basepauvreté<-rename.variable(basepauvreté,"Taux de pauvreté","Tx_pauv")
basepauvreté<-rename.variable(basepauvreté,"PIB/hab","PIB_hab")
basepauvreté<-rename.variable(basepauvreté,"Salaire moyen","Sal_moy")
basepauvreté<-rename.variable(basepauvreté,"Indice de Gini","Gini")
basepauvreté<-rename.variable(basepauvreté,"Indice de corruption","Corrup")
basepauvreté<-rename.variable(basepauvreté,"Taux de chômage","Tx_chom")
basepauvreté<-rename.variable(basepauvreté,"Dépenses en Santé","Sante")
basepauvreté<-rename.variable(basepauvreté,"Espérance de vie à la naissance","Espce_vie")
basepauvreté<-rename.variable(basepauvreté,"Dépenses intérieur brutes de R+D","RetD")
names(basepauvreté)

```

#On vérifie la nature des variables (vérifier qu'elles soient bien en mode quantitatives càd "numeric")

```
str(basepauvreté)
```

#On sort ensuite les graphiques de corrélation des variables actives en fonction du taux de pauvreté

```
library("tidyverse")
```

```
df <- basepauvreté
```

```

p <- ggplot(data = df,mapping=aes(PIB_hab,Tx_pauv)) + geom_point(mapping=aes(PIB_hab,Tx_pauv,
color=Continents), size=3) + geom_smooth(se=FALSE)
p + labs(title="Graphique de corrélation n°1",
caption="Graphique élaboré à partir de : https://donnees.banquemondiale.org/ et
https://data.oecd.org/fr/",
x="PIB par habitant",
y="Taux de pauvreté") +
theme_linedraw()

```

```

p1 <- ggplot(data = df,mapping=aes(Sal_moy,Tx_pauv)) + geom_point(mapping=aes(Sal_moy,Tx_pauv,
color=Continents), size=3) + geom_smooth(se=FALSE)
p1 + labs(title="Graphique de corrélation n°2",
caption="Graphique élaboré à partir de : https://donnees.banquemondiale.org/ et
https://data.oecd.org/fr/",
x="Salaire moyen",
y="Taux de pauvreté") +
theme_linedraw()

```

```

p2 <- ggplot(data = df,mapping=aes(Gini,Tx_pauv)) + geom_point(mapping=aes(Gini,Tx_pauv, color=Continents),
size=3) + geom_smooth(se=FALSE)
p2 + labs(title="Graphique de corrélation n°3",
caption="Graphique élaboré à partir de : https://donnees.banquemondiale.org/ et
https://data.oecd.org/fr/",
x="Indice de Gini",
y="Taux de pauvreté") +
theme_linedraw()

```

```

p3 <- ggplot(data = df,mapping=aes(Corrup,Tx_pauv)) + geom_point(mapping=aes(Corrup,Tx_pauv,
color=Continents), size=3) + geom_smooth(se=FALSE)
p3 + labs(title="Graphique de corrélation n°4",
caption="Graphique élaboré à partir de : https://donnees.banquemondiale.org/ et
https://data.oecd.org/fr/",
x="Indice de corruption",
y="Taux de pauvreté") +
theme_linedraw()

```

```

p4 <- ggplot(data = df, mapping=aes(IDH, Tx_pauv)) + geom_point(mapping=aes(IDH, Tx_pauv,
color=Continents), size=3) + geom_smooth(se=FALSE)
p4 + labs(title="Graphique de corrélation n°5",
caption="Graphique élaboré à partir de : https://donnees.banquemondiale.org/ et
https://data.oecd.org/fr/",
x="IDH",
y="Taux de pauvreté") +
theme_linedraw()

p5 <- ggplot(data = df, mapping=aes(Tx_chom, Tx_pauv)) + geom_point(mapping=aes(Tx_chom, Tx_pauv,
color=Continents), size=3) + geom_smooth(se=FALSE)
p5 + labs(title="Graphique de corrélation n°6",
caption="Graphique élaboré à partir de : https://donnees.banquemondiale.org/ et
https://data.oecd.org/fr/",
x="Taux de chômage",
y="Taux de pauvreté") +
theme_linedraw()

p6 <- ggplot(data = df, mapping=aes(Sante, Tx_pauv)) + geom_point(mapping=aes(Sante, Tx_pauv,
color=Continents), size=3) + geom_smooth(se=FALSE)
p6 + labs(title="Graphique de corrélation n°7",
caption="Graphique élaboré à partir de : https://donnees.banquemondiale.org/ et
https://data.oecd.org/fr/",
x="Dépenses en Santé",
y="Taux de pauvreté") +
theme_linedraw()

p7 <- ggplot(data = df, mapping=aes(Espce_vie, Tx_pauv)) + geom_point(mapping=aes(Espce_vie, Tx_pauv,
color=Continents), size=3) + geom_smooth(se=FALSE)
p7 + labs(title="Graphique de corrélation n°8",
caption="Graphique élaboré à partir de : https://donnees.banquemondiale.org/ et
https://data.oecd.org/fr/",
x="Espérance de vie",
y="Taux de pauvreté") +
theme_linedraw()

p8 <- ggplot(data = df, mapping=aes(RetD, Tx_pauv)) + geom_point(mapping=aes(RetD, Tx_pauv,
color=Continents), size=3) + geom_smooth(se=FALSE)
p8 + labs(title="Graphique de corrélation n°9",
caption="Graphique élaboré à partir de : https://donnees.banquemondiale.org/ et
https://data.oecd.org/fr/",
x="Dépenses en R&D",
y="Taux de pauvreté") +
theme_linedraw()

```

II) Analyse en Composantes Principales !!!!

A) Valeurs propres et nombres d'axes

Statistiques descriptives.

```

summary(basepauvreté[,3:12])
round(sapply(basepauvreté[,3:12], mean), 2)
round(sapply(basepauvreté[,3:12], sd), 2)

```

####Histogrammes des variables

```
library(RColorBrewer)
par(mfrow=c(4,3))
hist(basepauvreté$Tx_pauv, main="Histogramme Taux de pauvreté", xlab="Taux de pauvreté",
ylab="Fréquences", col=brewer.pal(n = 7, name = "YlOrRd"))
hist(basepauvreté$PIB_hab, main="Histogramme PIB par habitant", xlab="PIB par habitant", ylab="Fréquences",
col=brewer.pal(n = 6, name = "Greens"))
hist(basepauvreté$Sal_moy, main="Histogramme Salaire moyen", xlab="Salaire moyen", ylab="Fréquences",
col=brewer.pal(n = 7, name = "PuRd"))
hist(basepauvreté$Gini, main="Histogramme Indice de Gini", xlab="Indice de Gini", ylab="Fréquences",
col=brewer.pal(n = 8, name = "Reds"))
hist(basepauvreté$Corrup, main="Histogramme Taux de corruption", xlab="Taux de corruption",
ylab="Fréquences", col=brewer.pal(n = 8, name = "Purples"))
hist(basepauvreté$IDH, main="Histogramme IDH", xlab="IDH", ylab="Fréquences", col=brewer.pal(n = 6, name =
"Greys"))
hist(basepauvreté$Tx_chom, main="Histogramme Taux de chômage", xlab="Taux de chômage",
ylab="Fréquences", col=brewer.pal(n = 6, name = "Blues"))
hist(basepauvreté$Sante, main="Histogramme Dépenses en Santé", xlab="Dépenses en Santé",
ylab="Fréquences", col=brewer.pal(n = 8, name = "Oranges"))
hist(basepauvreté$Espce_vie, main="Histogramme Espérance de vie", xlab="Espérance de vie",
ylab="Fréquences", col=brewer.pal(n = 5, name = "GnBu"))
hist(basepauvreté$RetD, main="Histogramme Dépenses en R&D", xlab="Dépenses en R&D", ylab="Fréquences",
col=brewer.pal(n = 9, name = "Spectral"))
```

####Skewness et Kurtosis pour confirmer la distribution des histogrammes

```
library(moments)
round(sapply(basepauvreté[,3:12],skewness),2)
round(sapply(basepauvreté[,3:12],kurtosis),2)
```

####Matrice de corrélation

```
cor(basepauvreté[,c(3:12)])
round(cor(basepauvreté[,c(3:12)]),2)
library(corrplot)
cor1 <- cor(basepauvreté1[,c("Tx_pauv", "PIB_hab", "Sal_moy", "Gini", "Corrup", "IDH", "Tx_chom", "Sante",
"Espce_vie", "RetD")],use="complete.obs",method=c("spearman"))
cor1
col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA"))
corrplot(cor1, method="color", col=col(200),
type="upper",
addCoef.col = "black")
```

####Finalement on regarde s'il y a des pays atypiques

```
par(mfrow=c(3,4))
boxplot(`basepauvreté`$`Tx_pauv`, xlab="Taux de pauvreté", col="brown")
boxplot(`basepauvreté`$`PIB_hab`, xlab="PIB/hab", col="grey")
boxplot(`basepauvreté`$`Sal_moy`, xlab="Salaire moyen", col="grey")
boxplot(`basepauvreté`$`Gini`, xlab="Indice de Gini", col="grey")
boxplot(`basepauvreté`$`Corrup`, xlab="Indice de corruption", col="grey")
boxplot(`basepauvreté`$`IDH`, xlab="IDH", col="grey")
boxplot(`basepauvreté`$`Tx_chom`, xlab="Taux de chômage", col="grey")
boxplot(`basepauvreté`$`Sante`, xlab="Dépenses en Santé", col="grey")
boxplot(`basepauvreté`$`Espce_vie`, xlab="Esperance de vie à la naissance", col="grey")
boxplot(`basepauvreté`$`RetD`, xlab="Dépenses intérieur brutes de R&D", col="grey")
```



```

#### Valeurs propres et tableau d'inertie associé à chaque variable
library(FactoMineR)
library(factoextra)
library(corrplot)

# Création nouvelle base contenant que les variables actives
basepauvreté.actifs=basepauvreté[,c(4:12)]
basepauvreté.illus=basepauvreté[,3]
basepauvreté1=basepauvreté.actifs
View(basepauvreté1)

# Tableau d'inertie
res.pca=PCA(basepauvreté1)
round(res.pca$eig,2)
fviz_eig(res.pca, addlabels = TRUE)

# Valeurs propres
acp.basepauvreté=princomp(basepauvreté1, cor=T, scores=T)
val.propres=acp.basepauvreté$sdev^2
print(val.propres)

## B) Variables ; contributions, corrélations, cosinus carrés

#### Contributions
round(res.pca$var$contrib,2)
# Contributions des variables à l'axe 1
fviz_contrib(res.pca, choice = "var", axes = 1, col="black")
# Contributions des variables à l'axe 2
fviz_contrib(res.pca, choice = "var", axes = 2, col="black")
# Contributions des variables aux axes 1 et 2
fviz_contrib(res.pca, choice = "var", axes = 1:2, col="black")

## Corrélations
round(res.pca$var$cor,2)
res <- get_pca_var(res.pca)
corrplot(res$cor, is.corr=FALSE, method="circle", tl.srt=45, tl.col="#004400", col=brewer.pal(n=9,
name="RdYlBu"),
addCoef.col="black")

## Cosinus carrés
round(res.pca$var$cos2,2)
fviz_cos2(res.pca, choice = "var", axes = 1:2, col="black")

corrplot(res$cos2, is.corr=FALSE, method="circle", tl.srt=45, tl.col="#004400", col=brewer.pal(n=9,
name="PiYG"),
addCoef.col="black")

#Biplot
fviz_pca_biplot(res.pca, repel = TRUE,
  col.var = "#2E9FDF",
  col.ind = "red"
)

```

```

# Résumé des contributions pour construire le tableau récapitulatif des contrib positives, négatives de chaque
variable aux axes
dimdesc(res.pca)

## C) Cercle de corrélation

# Cercle de corrélation des variables actives
fviz_pca_var (res.pca, col.var = "cos2",
              gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
              repel = TRUE)

## D) Définition des variables latentes

# Description de la dimension 1
res.desc <- dimdesc(res.pca, axes = c(1,2), proba = 0.05)
res.desc$Dim.1

# Description de la dimension 2
res.desc$Dim.2

# Cercle de corrélation comprenant notre variable illustrative : le taux de pauvreté
res.pca1=PCA(basepauvreté[,3:12], quanti.sup=1)
fviz_pca_var (res.pca1, col.var = "cos2",
              gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
              repel = TRUE)

## E) Projection des individus sur le plan factoriel

## Contributions, coordonnées, cosinus carrés

###Contributions
round(res.pca$ind$contrib,2)
# Contributions des individus à l'axe 1
fviz_contrib(res.pca, choice = "ind", axes = 1, col="black", top=10)
# Contributions des individus à l'axe 2
fviz_contrib(res.pca, choice = "ind", axes = 2, col="black", top=10)

###Coordonnées
res <- get_pca_ind(res.pca)
round(res.pca$ind$coord,2)

###Cosinus carrés
round(res.pca$ind$cos2,2)
# Cosinus carré des individus à l'axe 1
fviz_cos2(res.pca, choice = "ind", axes = 1, col="black", top=10)
# Cosinus carré des individus à l'axe 2
fviz_cos2(res.pca, choice = "ind", axes = 2, col="black", top=10)

## Projection des individus sur les dimensions 1 et 2
fviz_pca_ind (res.pca, col.ind = "cos2",
              gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
              repel = TRUE)

## Afficher les 10 pays les plus contributifs

```

```

par(mfrow=c(1,1))
plot(res.pca,title="Les 10 pays les plus contributifs",select="contrib10",col.ind = "purple")

## Afficher les pays avec un cosinus supérieur à 0,9
plot(res.pca, title="Pays dont le cosinus carré est supérieur à 0,90", select="cos2 0.9", axes=c(1,2), col.ind ="red")

### III) Corrélation entre la variable à expliquer et les variables latentes !!!

## A) Régression linéaire multiple
lm=lm(basepauvreté$Tx_pauv~res.pca$ind$coord[,1]+res.pca$ind$coord[,2])
summary(lm)

# Analyse des résidus
basepauvreté$prediction=round(predict(lm),2)
basepauvreté$residu=round(residuals(lm),2)
basepauvreté$prediction
basepauvreté$residu

# Graphiques
library(gvlma)
par(mfrow=c(1,3))

boxplot(basepauvreté$residu, main="Boxplot des résidus", col="#FFCC99")
hist(basepauvreté$residu, xlab="Résidus", ylab="Nombre de pays", main="Histogramme des résidus",
freq=TRUE, col=brewer.pal(n = 8, name = "Reds"),cex.main=0.8)
qqnorm(basepauvreté$residu)
qqline(basepauvreté$residu, col="red")

# Coefficients d'asymétrie et d'aplatissement des résidus
library(moments)
round(skewness(basepauvreté$residu),3)
round(kurtosis(basepauvreté$residu),3)

# Qualité du modèle
library(gvlma)
gvlma(lm)

# Test paramétrique KS
ks.test(basepauvreté$residu, "pnorm", mean(basepauvreté$residu), sd(basepauvreté$residu))

# Test non paramétrique de Levene
library(car)
xA=c(1:29)
xB=c(30:59)
nA=length(xA)
nB=length(xB)
y=c(xA,xB)
groupe=as.factor(c(rep("A",nA),rep("B",nB)))
leveneTest(y,groupe)

## B) Relation entre la variable à expliquer observée et prédite par RLM

### Valeurs prédites par le modèle vs. valeurs réelles (+graphique)
basepauvreté$predict=predict(lm)
round(basepauvreté$predict,2)

```

```
basepauvreté$Tx_pauv  
basepauvreté$residu
```

```
par(mfrow=c(1,1))  
plot(basepauvreté$predict,basepauvreté$Tx_pauv, col=0, main = "Taux de pauvreté prédit par le modèle vs. taux observé", xlab="Taux de pauvreté prédit", ylab="Taux de pauvreté observé", abline(a=0, b=1, col="red", lty=1))  
text(basepauvreté$predict,basepauvreté$Tx_pauv,row.names(basepauvreté),cex=.8)
```

C) Régression logistique binaire

##Codage de Y par rapport à sa moyenne + model GLM

```
basepauvreté$Tx_pauvbis[basepauvreté$Tx_pauv<=15.12]="0"  
basepauvreté$Tx_pauvbis[basepauvreté$Tx_pauv>15.12]="1"  
basepauvreté$Tx_pauvbis
```

```
basepauvreté$Tx_pauvbisN=as.numeric(basepauvreté$Tx_pauvbis)  
modele=glm(basepauvreté$Tx_pauvbisN~res.pca$ind$coord[,1]+res.pca$ind$coord[,2],family=binomial)  
summary(modele)
```

##Codage de Y par rapport à sa médiane pcq l'axe 2 n'est significatif qu'au seuil de 5% (pas de 1%)

```
basepauvreté$Tx_pauvter[basepauvreté$Tx_pauv<=14.4]="0"  
basepauvreté$Tx_pauvter[basepauvreté$Tx_pauv>14.4]="1"  
basepauvreté$Tx_pauvter
```

```
basepauvreté$Tx_pauvterN=as.numeric(basepauvreté$Tx_pauvter)  
modele1=glm(basepauvreté$Tx_pauvterN~res.pca$ind$coord[,1]+res.pca$ind$coord[,2],family=binomial)  
summary(modele1)
```

##Recherche du seuil pour axe 2 significatif même à 0,1 %

```
basepauvreté$Tx_pauvqua[basepauvreté$Tx_pauv<=13.3]="0"  
basepauvreté$Tx_pauvqua[basepauvreté$Tx_pauv>13.3]="1"  
basepauvreté$Tx_pauvqua  
basepauvreté$Tx_pauvquaN=as.numeric(basepauvreté$Tx_pauvqua)  
modele2=glm(basepauvreté$Tx_pauvquaN~res.pca$ind$coord[,1]+res.pca$ind$coord[,2],family=binomial)  
summary(modele2)
```

Partie 7 : Bibliographie

Alvaredo.F, Chancel.L, Piketty.T, Saez.E, Zucman.G, “Le Rapport sur les inégalités mondiales 2018”, décembre 2017

<http://ses.ens-lyon.fr/actualites/rapports-etudes-et-4-pages/rapport-sur-les-inegalites-mondiales-2018-wid-world-decembre-2017> (consulté le 26/09/2019)

Blampain.N, “L’espérance de vie par niveau de vie : chez les hommes, 13 ans d’écart entre les plus aisés et les plus modestes”, *Division Enquêtes et études démographiques, Insee*, 2018

<https://www.insee.fr/fr/statistiques/3319895> (consulté le 22/10/2019)

Café Pédagogique consultable en ligne :

<http://www.cafepedagogique.net/lemensuel/lenseignant/schumaines/ses/Pages/2008/Sant%C3%A9etPauvret%C3%A9danslespaysend%C3%A9veloppement.aspx> (consulté le 22/10/2019)

CARTO “Les effets de la crise sur la société : la Pauvreté dans le Monde”, *Le Monde en Carte*, juin 2012, numéro 11, p12-22.

<http://www.sciencespo.fr/ceri/fr/carto/dossier-carto-la-pauvrete-dans-le-monde> (consulté le 24/09/2019)

Clément.M “Mieux comprendre les facteurs de risque de pauvreté en conditions de vie en contrôlant les caractéristiques inobservées fixes”, *Insee – DSDS – Unité des études démographiques et sociales*, p.57 (consulté le 20/10)

<https://www.insee.fr > statistiques > fichier>

Droz-Georget.C “PIB par habitant : quelle tendance en Europe et en France ?”, *Office fédéral de la statistique*, avril 2019

<https://www.ifrap.org/etat-et-collectivites/pib-par-habitant-quelle-tendance-en-europe-et-en-france> (consulté le 09/10/2019)

Förster.M et Pearson.M, “Distribution des revenus et pauvreté dans la zone de l’OCDE : tendances et déterminants”, *Revue économique de l’OCDE*, 2002 (numéro 34), pages 7 à 41

<https://www.cairn.info/revue-economique-de-l-ocde-2002-1-page-7.htm> (consulté le 23/10/2019)

INSEE, définition du taux de pauvreté, consultable en ligne

<https://www.insee.fr/fr/metadonnees/definition/c1320> (consulté le 29/09/2019)

INSEE, définition du taux de chômage, consultable en ligne :

<https://www.insee.fr/fr/metadonnees/definition/c1129> (consulté le 17/10/2019)

INSEE, “Dépense intérieure de recherche et développement / DIRD”, *INSEE*, 2019.

<https://www.insee.fr/fr/metadonnees/definition/c1226> (consulté le 28/10/2019)

INSEE, “Dépenses extérieures de recherche et développement des entreprises (DERDE)”, *Enseignement supérieur - recherche*, 2002.

<https://www.enseignementsup-recherche.gouv.fr/reperes/telechar/res/res00/rapmeto00.pdf>
(consulté le 28/10/2019)

Jahan.S, “Rapports sur le développement humain”, publié par *le Programme des Nations Unies pour le développement* (PNUD), 2017
<https://www.undp.org/content/undp/fr/home/librarypage/hdr/> (consulté le 15/10/2019)

La Banque Mondiale consultable en ligne :
<https://donnees.banquemondiale.org/> (consulté le 22/10/2019)

La finance pour tous, “PIB (produit intérieur brut) définition et enjeux”, mise à jour le 5 avril 2019.
<https://www.lafinancepourtous.com/decryptages/politiques-economiques/theories-economiques/pib/>

Observatoire des inégalités, “Les seuils de pauvreté en France” 10 septembre 2019.
https://www.inegalites.fr/Les-seuils-de-pauvrete-en-France?id_theme=15

OCDE consultable en ligne :
<https://data.oecd.org/fr/> (consulté le 22/10/2019)

OCDE, définition salaire moyen, consultable en ligne :
<https://data.oecd.org/fr/earnwage/salaires-moyens.htm> (consulté le 15/10/2019)

OCDE, définition inégalité de revenus, consultable en ligne :
<https://data.oecd.org/fr/inequality/inegalite-de-revenu.htm> (consulté le 13/10/2019)

SES.WEBCLASS, “indice de développement humain IDH”, *Chapitre 1. Quelles sont les sources de la croissance économique ?*, SES.WEBCLASS consultable en ligne :
<http://ses.webclass.fr/notion/indice-developpement-humain-idh> (consulté le 17/10/2019)

Transparency international France, “Indices de perception de la corruption de Transparency International”, 2016.
<https://transparency-france.org/publications/indices-de-perception-de-corruption/#.Xd5Ez-TsbIU>

Notre base de données finale (au format CSV), *Google Drive* :
https://docs.google.com/spreadsheets/d/e/2PACX-1vRs8JjiBAAeFhiQgClMbc6iMVn88lh14LjPz4jxAt eJNLQxnAy748VRHGuu_vBft1IiMID_ARo1K6i/pub?output=xlsx

Partie 8 : Table des matières

Partie 1 : Introduction	2
Partie 2 : Analyse détaillée du sujet	5
Une variable illustrative	5
Des variables actives	7
1- Pib par habitant (X_1) et salaire moyen (X_2)	8
2- Indice de Gini (X_3)	10
3- Indice de corruption (X_4)	12
4- Indice de développement humain (X_5)	13
5- Taux de chômage (X_6)	15
6- Dépenses de Santé (X_7) et espérance de vie à la naissance (X_8)	16
7- Dépenses intérieures brutes de recherche et développement (X_9)	19
Partie 3 : Analyse en Composantes Principales (ACP)	23
Valeurs propres et nombre d'axes	23
1- Analyse descriptive	23
2- Valeurs propres et tableau d'inertie	29
Variables ; contributions, corrélations, cosinus carrés	31
1- Contributions	31
2- Corrélations	32
3- Cosinus carrés	33
Cercle des corrélations	35
Définition des variables latentes	37
1- Interprétation de la composante n°1	37
2- Interprétation de la composante n°2	38
3- Ajout de la variable illustrative à l'ACP	39
Projection des individus sur le plan factoriel, interprétation	40
1- Projection des individus sur les dimensions 1 et 2	41
→ Analyse des pays situés aux extrémités des axes	42
2- Identification des individus les plus importants dans l'explication des 2 axes	43
Partie 4 : Corrélation entre la variable à expliquer et les variables latentes	46
Régression linéaire multiple	46
1- Estimation des coefficients	46
2- Analyse des résidus	47
Relation entre la variable à expliquer observée et prédite par RLM	51
Partie 5 : Conclusion et discussion des résultats	56
Partie 6 : Annexes	57
Tableaux supplémentaire utilisés dans l'analyse	57
Script R studio	61
Partie 7 : Bibliographie	68

